Characterizing the Genomic Diversity, Evolution and Phylogeography of Respiratory Syncytial Virus Genotype ON1 in Kenya

Thesis

CHARACTERIZING THE GENOMIC DIVERSITY, EVOLUTION AND PHYLOGEOGRAPHY OF RESPIRATORY SYNCYTIAL VIRUS GENOTYPE ON1 IN KENYA

Thesis submitted for award of degree of Doctor of Philosophy

Open University (UK)

Affiliated Research Centre

KEMRI-Wellcome Trust Research Programme, Kilifi, Kenya

James Richard Otieno

BSc Biochemistry, Moi University (Kenya)

MSc Bioinformatics, University of Leicester (UK)

Submission date
February 2019

KEMRI | Wellcome Trust

The Open University

DECLARATION

This thesis describes my work that was undertaken at the KEMRI-Wellcome Trust Research Programme under the supervision of Prof. James Nokes, Prof. Philippe Lemey and Dr. Charles Agoti in fulfilment of the requirements for the degree of Doctor of Philosophy at the Open University (UK). This dissertation is the result of my own work and contains nothing that is the outcome of work done in collaboration, except where specifically indicated in the text. The work described here has not been submitted for any other qualification at any other university or institution.

Sample collection and consenting was conducted as part of larger integrated studies on the surveillance of respiratory viruses across Kenya. Sequence data was generated at the Wellcome-Trust Sanger Institute (UK) and the KEMRI-Wellcome Trust Research Programme (Kilifi, Kenya)

James Richard Otieno

Kilifi (Kenya), February 2019

ABSTRACT

Background

In December 2010, a new genotype of respiratory syncytial virus (RSV) with a 72-nucleotide duplication within the attachment (G) gene was identified in Ontario, Canada, and named ON1. Using the ON1 as a unique tag, this study aimed to understand; (1) how new RSV variants are introduced, spread and persist in communities, (2) the genomic signatures that define the emergent RSV variants and whether such substitutions may be associated with potential fitness advantages, and (3) the patterns of RSV spread across geographically defined regions (local and global).

Methods

Partial G gene (n=483) and whole genome (n=184) sequence datasets collected between 2010 and 2016 were analyzed using genetic diversity, phylogenetics and statistical methods to understand the molecular epidemiology of RSV in Kilifi County, Coastal Kenya. Further, Kenyan (partial G gene; n=2526) and global (full G gene, n=2238; whole genome, n=1194) sequence datasets collected between 1977 to 2016 were analysed in a Bayesian framework for the inference of the phylogeographic history of local and global RSV spread, respectively.

Results

Following initial detection of the genotype ON1 in Kilifi in 2012, there was rapid replacement of the previously circulating RSV group A genotype GA2 by ON1 in subsequent epidemics. While this suggests elevated fitness of ON1 viruses, there was no clear evidence of altered pathogenicity of ON1 relative to GA2 in Kilifi. Signature

amino acid substitutions were identified between surface proteins (G, F), polymerase (L) and matrix M2-1 proteins of Kilifi ON1 and GA2 viruses, suggesting co-evolution amongst antigenic and non-antigenic genes of RSV variants. Genetic and phylogenetic analyses reaffirmed previous conclusions that each RSV epidemic is characterized by the frequent introduction of multiple variants, few of which persist across epidemics. Finally, the phylogeographic analyses predicted the northern hemisphere to be the major source population of RSV into the tropics and the southern hemisphere and virus spread between locations in close proximity to be important for virus persistence within a country.

Conclusions

Tracking the ON1 tag offered important insights into RSV evolution and transmission. The use of whole genome sequencing and surveying all the variation throughout the genome will be crucial for greater understanding, and potentially improved control, of this important pathogen. However, there is a need for a more targeted approach to RSV surveillance and sequencing that will help build a better picture of RSV spread at different scales.

DEDICATION

This thesis is dedicated to:
⇐family ⇒
Rehema Luvuno Chimera and Liam Jared Otieno

⇐parents ⇒
Jared Aol Otieno and Rose Akinyi Yugi

⇐champions ⇒
Peter Ochieng Yugi (1970-2005) and Richard Otieno Snr (1928-2018)

I hope I made you all proud!

ACKNOWLEDGEMENTS

Words are never enough, but all I've got at this moment is a blank page and a keyboard. Perhaps a 21-gun parade, biggest hugs, or grandest meals would be more apt…but oh well;

"Thank You!"

To the Supervisors; James, Philippe and Charles, Ahsanteni Sana!

A special thank you to James who started this off by asking me on the 10th of September 2013 if I wanted to do a PhD. You remained unflinchingly positive, understanding and encouraging throughout. Hoorah!

To Philippe, many thanks for accepting to supervise a student from Kenya on that 12th day of January 2015. January is indeed a good month, in fact the greatest month; I know this as I was born in January! I have learnt a lot from you, and more yet to learn

Office 817: The friendship and laughter kept me going. I hope I can give back in the PhD after life.

VEC group members (including current, past and ex-officio) are an amazing and helpful lot. A special mention of Everlyn, Grieven, John Oketch, Chapa, Adema, Alex, Ann, Lewa, Martin, Sonal, Patrick, Sande, George G, Nelson and Matt Cotten.

IDeAL Team: Sam, Liz, Rita and Florence. Many thanks for the financial support and ensuring my records were up to date.

Family: Rehema and Liam for enduring the frequent absence and giving me a reason to soldier on. Parents and siblings for the immense pressure to be the first PhD in the village, the things you do for family!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Aa	Amino acid
ALRTI	Acute lower respiratory tract infections
ARI	Acute respiratory infection
BEAST	Bayesian Evolutionary Analysis and Sampling of Trees
BF	Bayes factor
Bp	Base pairs
BSSVS	Bayesian stochastic search variable selection
CDC-K	Centre for disease control - Kenya
CDS	Coding sequence
CGMR-C	Centre for geographic medicine research – coast
CSC	Centre scientific committee
Ct	Cycle threshold
CTMC	Continuous time markov chain
HCoV	Human Coronavirus
F	Fusion glycoprotein
FUBAR	Fast unconstrained bayesian approximation
FTD	Fast track diagnostics
G	Attachment glycoprotein
GLEAM	Global epidemic and mobility
GLM	Generalized linear model
HMPV	Human metapneumovirus
HKY	Hasegawa Kishino and Yano
HRV	Human Rhinovirus
HyPhy	Hypothesis testing using phylogenies
IFAT	Immunofluorescent antibody test
ILI	Influenza-like illness
IP	Inpatient
KCH	Kilifi County Hospital
KEMRI	Kenya Medical Research Institute
KHDSS	Kilifi Health and Demographic Surveillance System
KMoD	Kenya ministry of state for defence

KWTRP	KEMRI-Wellcome Trust Research Programme
L	RNA-dependent RNA polymerase
LRTI	Lower respiratory tract infections
M	Matrix protein
M2-1	Transcription and antitermination factor
M2-2	Transcription regulation and RNA replication
MCMC	Markov Chain Monte Carlo
MEME	Mixed effects model of evolution
ML	Maximum Likelihood
MAFFT	Multiple alignment using fast fourier transform
N	Nucleoprotein
NGS	Next generation sequencing
NP	Nasopharyngeal
NS1	Non-structural protein 1
NS2	Non-structural protein 2
nt	Nucleotide
OP	Oropharyngeal
ORF	Open reading frame
P	Phosphoprotein
PCR	Polymerase chain reaction
phyphy	Python HyPhy
PoPros	Potential primer sequences
RNA	Ribonucleic acid
RSV	Respiratory syncytial virus
RT- PCR	Reverse transcriptase - polymerase chain reaction
SARI	Severe acute respiratory syndrome
SERU	Scientific and ethics review unit
SH	Small hydrophobic protein
SLAC	Single likelihood ancestral counting
SNP	Single nucleotide polymorphism
SPReD	Studies of the pathways of transmission of respiratory virus disease
Tm	Melting temperature
TMRCA	Time to the most recent common ancestor

UK	United Kingdom
URTI	Upper respiratory tract infection
USA	United States of America
USAMRU-K	US army medical research directorate – Kenya
VEC	Viral Epidemiology and Control
WGS	Whole genome sequencing
WHO	World Health Organization
WTSI	Wellcome Trust Sanger Institute

CHAPTER ONE

Introduction

1.1 RSV Disease Burden

Human respiratory syncytial virus (RSV) is the most important cause of viral acute lower respiratory tract infections (ALRTI) in children worldwide and a health concern in terms of morbidity, mortality and costs (Simoes 1999; Cane 2001). The virus was first isolated from a captive chimpanzee in 1955 (Morris, Blount and Savage 1956) and thereafter identified to be a major human paediatric respiratory pathogen (Chanock and Finberg 1957; Chanock, Roizman and Myers 1957). Almost all individuals experience the first RSV infection by the age of two years, and about 1% of infants in their first year of life require hospitalization due to RSV-associated pneumonia or bronchiolitis (Henderson *et al.* 1979; Glezen *et al.* 1986; Nokes *et al.* 2004). Severe RSV infection early in life has been associated with later development of asthma and wheeze (Blanken *et al.* 2013). In addition, RSV is also an important cause of morbidity and mortality in the elderly and in adults with cardiopulmonary disease or with an impaired immune system (Falsey *et al.* 2005).

A recent global study (2017) by *Shi et al*. reported that there were as many as 33·1 million (uncertainty range [UR] 21·6–50·3) episodes of RSV-ALRI in 2015, and these resulted in about 3·2 million (2·7–3·8) hospital admissions and 59,600 (48,000–74,500) in-hospital deaths in children younger than 5 years (Shi *et al.* 2017). In fact, they estimated that the overall RSV-ALRI mortality could be as high as 118,200 (UR 94,600–149,400) if including deaths outside the hospital setting. Most of the RSV-associated burden is in the developing countries (Weber, Mulholland and Greenwood 1998; Williams *et al.* 2002; Nokes 2007; Nair *et al.* 2010; Shi *et al.* 2017). In Kenya,

about 85,000 RSV-associated infant severe lower respiratory tract infections (LRTI) cases have been estimated to occur per year (Nokes *et al.* 2008). Immunity to infection is not solid and repeat infection occurs throughout life (Connors *et al.* 1991; Hall *et al.* 1991; Falsey 2007; Agoti *et al.* 2012). Despite considerable effort to develop an RSV vaccine, none has been licensed so far (Neuzil 2016). Infants and young children at high risk for severe RSV disease can be substantially protected by the passive administration of a commercially available RSV-neutralizing monoclonal antibody (MAb), palivizumab (The IMpact-RSV Study Group 1998). An antiviral therapy, ribavirin, is available even though its efficacy is marginal and it is not recommended for routine use (Borkje 1982). The phenomenon of reinfection and difficulty in developing a vaccine may in part be due to the virus' antigenic diversity (Melero and Moore 2013). However, there is a diverse set of vaccines that are in the development pipeline and targeting young children, older adults, and pregnant women (Neuzil 2016; PATH 2019).

.

1.2 RSV Genome and Diversity

RSV is an RNA virus with an ~15.2 kb negative sense, single stranded genome made up of 10 genes encoding 11 proteins: the non-structural proteins (NS1, NS2) the nucleocapsid-associated proteins (N, P, L, M2-1, M2-2), matrix protein (M1) and surface glycoproteins (SH, G, F) (Cane 2001). A schematic representation of the RSV genome with the respective known function for each encoded protein is shown in, *Figure 1.1* (Espinoza *et al.* 2014).

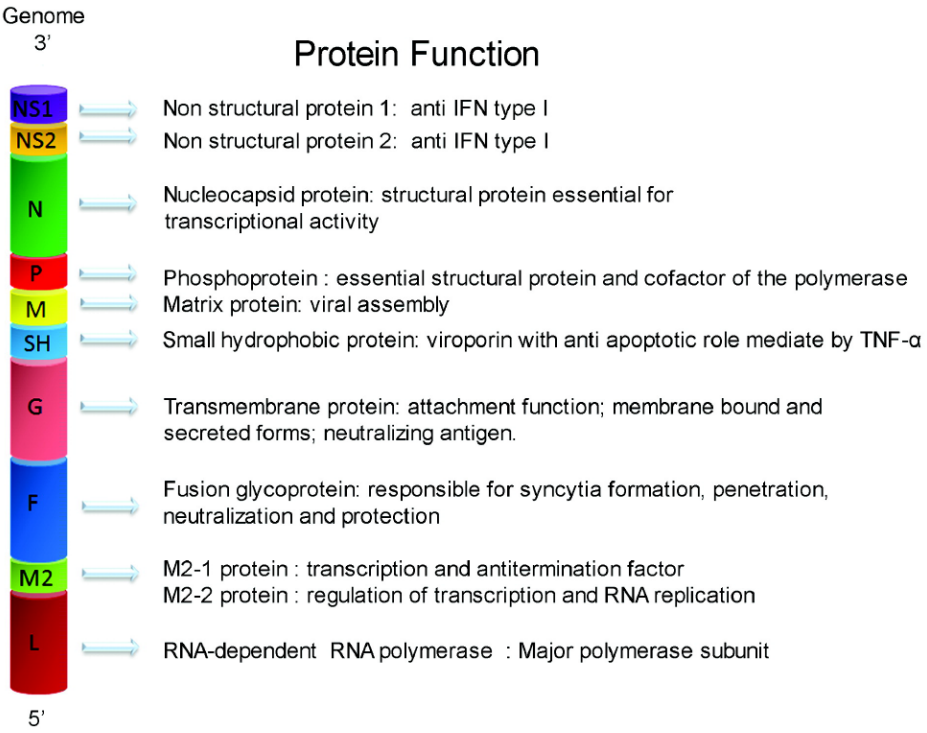


Figure 1.1: A schematic representation of the RSV genome indicating the known function for each encoded protein
(Espinoza *et al.* 2014)

A central characteristic of the majority of RNA virus populations is that they are highly dynamic, and this arises from short generation times (the average time taken for a viral genome or infected cell to produce another of its kind in a replication cycle), large population sizes, and high mutation frequencies (Rodrigo 1999; Belshaw *et al.* 2008; Peck and Lauring 2018). The accumulation of diversity often arises during genome replication whereby ~1 mutation per genome is generated as a result of an error-prone viral RNA-dependent RNA polymerase that lacks proofreading mechanisms (Holland *et al.* 1982; Steinhauer, Domingo and Holland 1992). During an infection, as a consequence of the high mutation rates, a heterogeneous population of closely related mutants is produced that is commonly referred to as a "quasispecies" (Domingo, Sheldon and Perales 2012). While it is thought that the

diversity in RSV is primarily driven by the error-prone polymerase, additional mechanisms of generation of diversity in other viruses include copy-choice recombination (the viral polymerase switches templates during replication) and reassortment (exchange of genetic segments during coinfection in segmented viruses).

RSV strains are classified into two groups (A and B), *Figure 1.2*, based on antigenic and genetic variability in some of the structural proteins (Anderson *et al.* 1985; Mufson *et al.* 1985; Cristina *et al.* 1990). Phylogenetic studies of RSV molecular epidemiology have primarily focused on G-gene sequences. These studies have identified multiple distinct viruses within the two groups called genotypes (Peret *et al.* 1998, 2000; Cane 2001). Peret *et al* in 1998 defined RSV genotypes based on sequence clusters in phylogenetic trees. Genotypes GA1 to GA5 were identified for RSV-A, resulting in amino acid level intergenotypic differences ranging between 10-28%. For RSV-B, genotypes GB1 to GB4 were identified and the resultant intergenotypic differences ranged from 7-19% at the amino acid level (Peret *et al.* 1998). The intergenotypic differences were calculated based on the G-protein's 270nt second hypervariable region. The number of genotypes subsequently expanded with the rise, identification and classification of more circulating variants (Venter *et al.* 2001; Venter, Collinson and Schoub 2002; Trento *et al.* 2003; Zlateva *et al.* 2004, 2005; Eshaghi *et al.* 2012). These genotypes have informed current understanding of RSV molecular epidemiology, and it is thought that genotype genetic variability and replacements contribute to enabling the virus to cause yearly outbreaks (Peret *et al.* 2000; Yamaguchi *et al.* 2011).

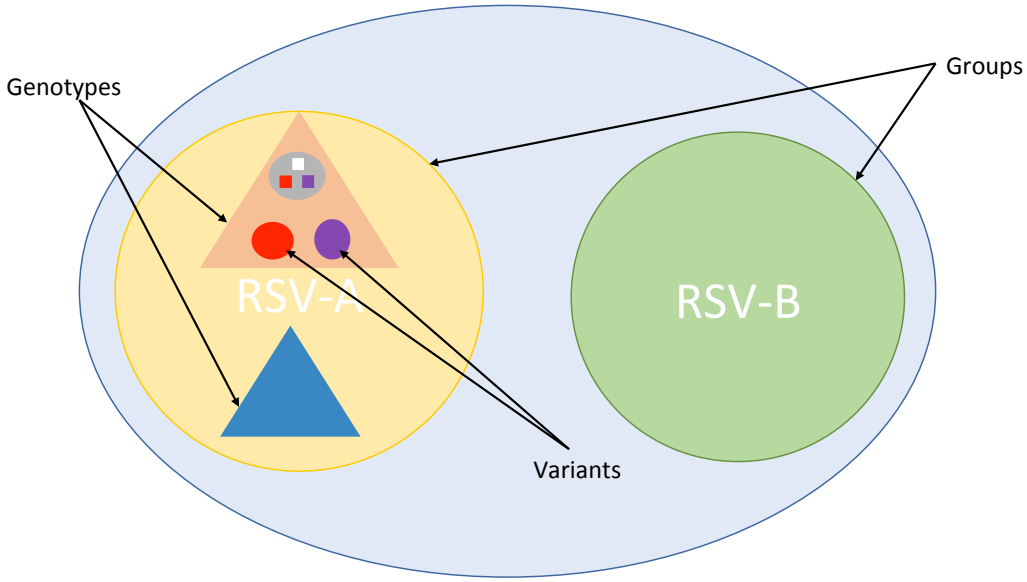


Figure 1.2: A schematic representation of the RSV classification into groups, genotypes and variants

1.3 RSV Epidemiology

RSV infections and disease in most places exhibit clear seasonality. Annual, biannual and biennial epidemics have been observed in different parts of the world (Weber, Mulholland and Greenwood 1998; Centers for Disease Control and Prevention (CDC) 2004; Mlinaric-Galinovic *et al.* 2012; Agoti *et al.* 2014a). The two RSV groups often co-circulate in epidemics with RSV-A generally occurring more frequently (Michael Hendry *et al.* 1986; Hendry, Pierik and McIntosh 1989; Hall *et al.* 1990; Rajala *et al.* 2003; Fodha *et al.* 2004). The pattern of alternation in dominance between RSV-A and B differs from place to place (Waris 1991; Cane 2001). In addition, epidemics are usually comprised of multiple genotypes even though the proportions of each genotype varies and genotype replacements are observed over time (Peret *et al.* 1998). The role and magnitude of genotype-specific herd immunity in driving the alternating patterns of change is not clear (Sande *et al.* 2013). Further, no conclusive association has been made between RSV groups or genotypes and disease severity based on the

G-gene studies (Walsh *et al.* 1991; Panayiotou *et al.* 2014; Yoshihara *et al.* 2016; Otieno *et al.* 2017). While the existence of such an association would inform control strategies on prioritization of strains that need most urgent attention, little is known about the interaction between the non-structural proteins and disease severity or strain persistence in communities.

An RSV genotype can be further divided into variants, *Figure 1.2*, which can either be (i) imported variants that show greater genetic divergence than expected from *in situ* diversification (Agoti *et al.* 2015b; Otieno *et al.* 2016), or (ii) local variants arising from recent introduction which subsequently diversify *in situ* (without time for purifying selection from, for example inter-epidemic bottlenecks) (Agoti *et al.* 2017). It has previously been shown that within RSV epidemics, there is co-circulation of viruses belonging to different RSV groups, genotypes and variants both imported and local, (Agoti *et al.* 2015b, 2017; Otieno *et al.* 2016).

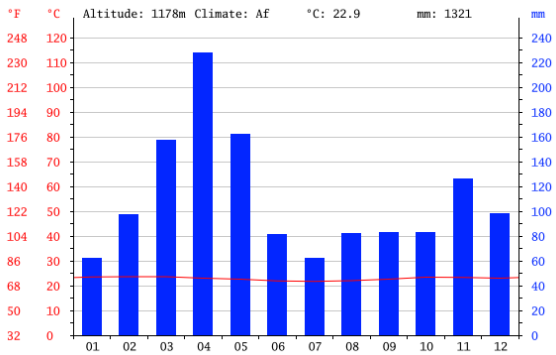
Generally, RSV outbreaks occur in the late fall and winter in the temperate regions and during the rainy seasons in the tropical regions (Moura *et al.* 2006; Goddard *et al.* 2007; Meerhoff *et al.* 2009; Murray *et al.* 2012; Obando-Pacheco *et al.* 2018). However, associations between RSV and weather vary across years and geographic locations (Haynes *et al.* 2013). Most RSV seasons last between 5-6 months, with shorter seasons of 3-4months and longer seasons of up to 10 months reported for some countries (Obando-Pacheco *et al.* 2018). RSV seasonality is fairly consistent within most regions from year to year, with minor variations of 1-3 weeks in the start, end and/or peak of RSV activity. For countries with large territories and different

regional climatic regions such as Brazil, USA or Australia, intra-country differences in seasonality have been reported (Obando-Pacheco *et al.* 2018).

1.3.1 ***RSV Epidemiology in Kenya***

Kilifi County, located in the South Eastern coastal Kenya, is characterized by a bimodal rainfall pattern with long rains during March-May and short rains in October-November (http://en.climate-data.org/location/11152/ accessed on 30/03/2015); *Figure 1.3A*. Kisumu, located in the Western part of Kenya, equally has two rainy seasons from March through June and November through December albeit with higher annual and monthly average rainfall than Kilifi (http://en.climate-data.org/location/715071/ accessed on 30/03/2015); *Figure 1.3B*. Annual RSV epidemics in Kilifi begin late October peaking from January to March, *Figure 1.4*. In western Kenya, however, RSV annual epidemics peak between April and July (*Figure 1.5*), somewhat coincidental with the rainy season (Emukule *et al.* 2014). Therefore, RSV epidemics in these two regions of Kenya are seemingly and interestingly out of synchrony by about 3-4 months.

A: Kisumu



B: Kilifi

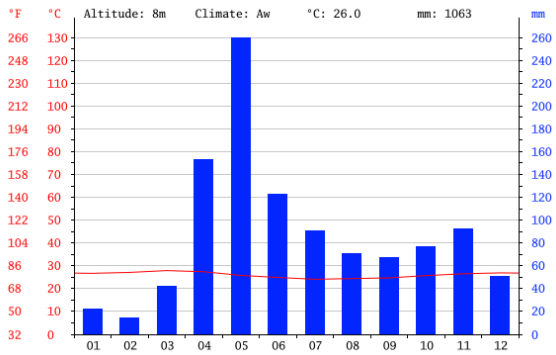


Figure 1.3: Monthly avaerage rainfall and temperature in Kisumu and Kilifi, Kenya

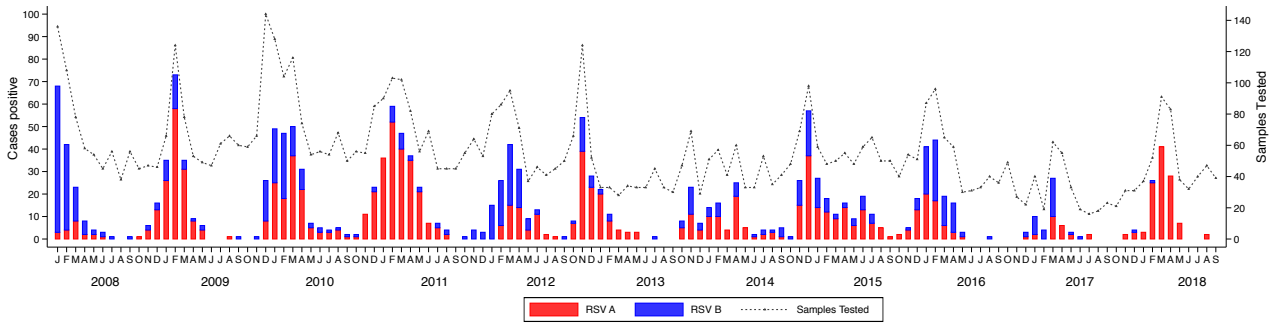


Figure 1.4: Seasonality of RSV in Kilifi, Coastal Kenya, 2008– 2018

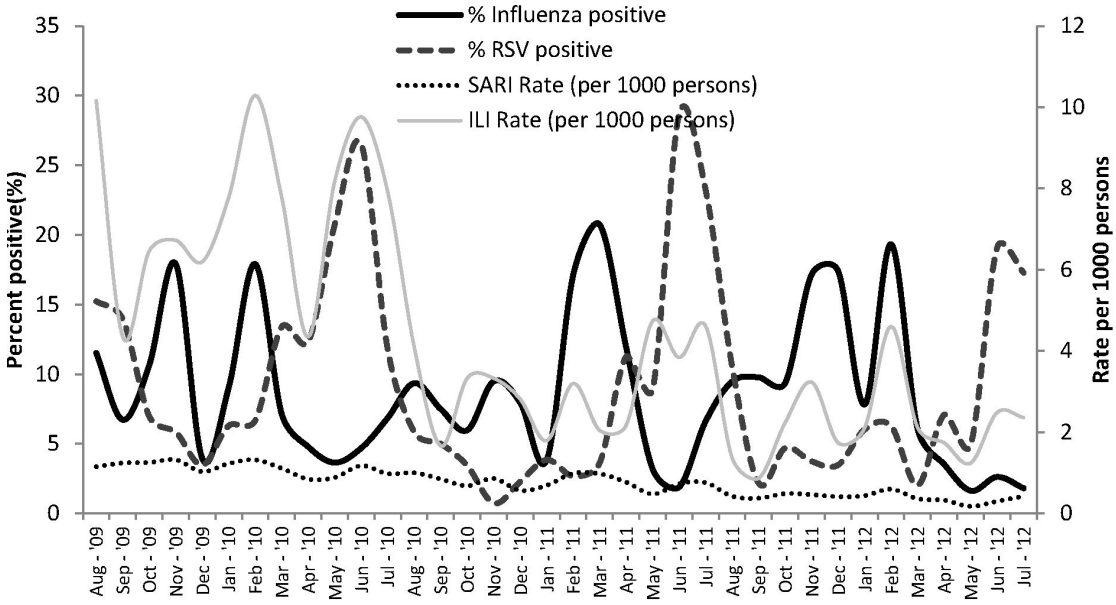


Figure 1.5: Seasonality of RSV and other viruses in Western Kenya, 2009– 2012 (Emukule *et al.* 2014)

Understanding the mechanisms underlying the seasonal patterns of RSV remains a challenge and cannot be ascribed to climate alone (White *et al.* 2007). While prior infection can modulate disease severity, immunity to RSV infection is short-lived resulting in relatively common reinfections (Melero *et al.* 1997). At the population level, variation in herd immunity and socio-demographic factors between different communities are potential determinants of RSV infection transmission (Anderson *et al.* 1991; Peret *et al.* 2000; White *et al.* 2005; Zlateva *et al.* 2007; Lemey *et al.* 2014).

Factors such as HIV prevalence and malnutrition have also been found to potentially affect disease epidemiology in sub-Saharan Africa (Madhi *et al.* 2000; Preidis *et al.* 2011). Therefore, an interplay of multiple factors determines the mechanics of virus transmission, i.e. introduction into a particular location, spread (or not) within and between locations, persistence or fade-out between epidemics; *Figure 1.6*. It follows that a study of local RSV transmission patterns in Kenya may not only elucidate the nature of RSV spread within and between the different parts of the country but also the potential factors influencing the countrywide disease transmission process (Out of Africa 2014).

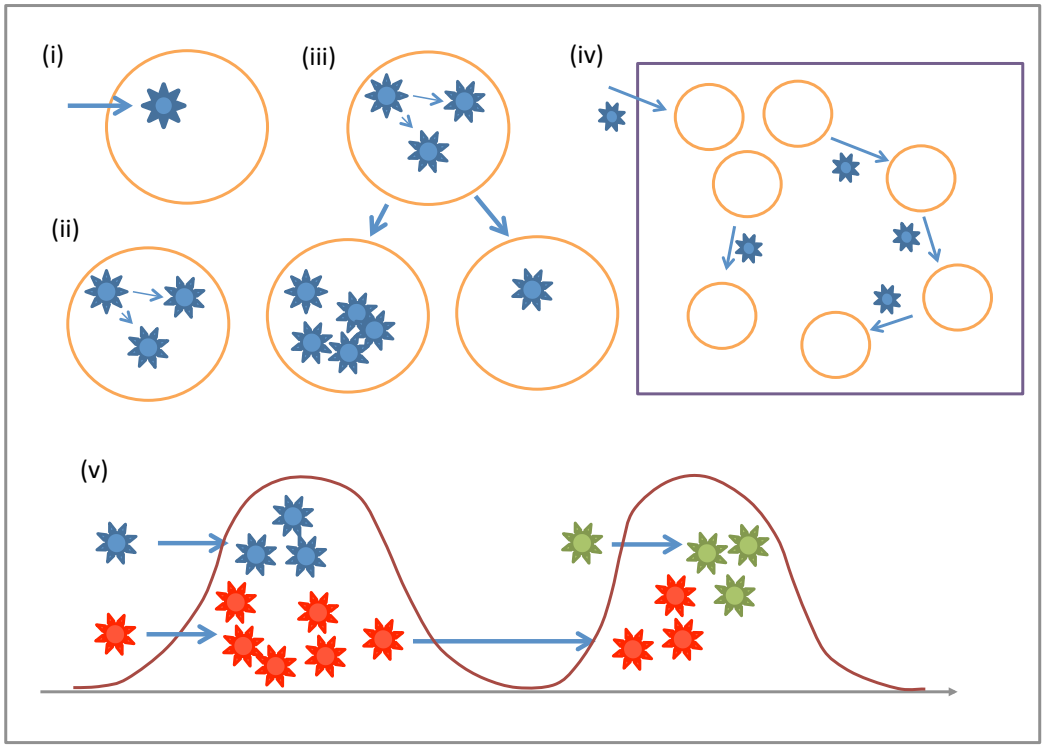


Figure 1.6: Virus strain introduction, spread, fadeout and persistence schema.
A virus (star shape) may be introduced to a location (circle) and (i) not spread, or spread (ii) locally but not to other locations, (iii) between locations with various levels of local transmission, or (iv) countrywide (square) from a common source or in some order that may or may not be predictable, and (v) may (red virus) or may not (blue virus) persist between seasonal outbreaks. Here the term virus strain is interchangeable with variant, lineage, genotype, or group or species.

1.4 Immunity to RSV

RSV repeatedly re-infects throughout life and re-infections can be severe (Nokes *et al.* 2008). Re-infection suggests an absence of sterilizing immunity to infection and disease and may in part relate to genetic diversity of the virus (Cane, Matthews and Pringle 1991; Sullender *et al.* 1991). Protective immunity in both infants and elderly adults correlates strongly with host factors such as the duration of circulating RSV neutralizing antibodies, the role of mucosal IgA and RSV-specific resident memory T cells (Johnson *et al.* 1987; Hendry *et al.* 1988; Habibi *et al.* 2015; Jozwik *et al.* 2015).

The surface of an infectious RSV virion contains 3 virus encoded proteins; F, G and SH that are exposed to the host immune system. Studies with monoclonal antibodies have confirmed that two of these surface proteins, F and G, are the targets of antibody responses (Cote *et al.* 1981). Three distinct epitope regions have been identified on the G protein mainly through reactivity with murine monoclonal antibodies, i.e. conserved, group-specific and strain-specific epitopes (Martínez, Dopazo and Melero 1997; Polack *et al.* 2005; Chirkova *et al.* 2013). Two forms of the G protein are synthesised in the course of infection; a membrane anchored form and a 6-9kD smaller soluble form (Hendricks *et al.* 1987). The soluble form is translated through an alternative in-frame start codon and is thought to act as an antigenic decoy by binding to host antibody and thus helping the virus to evade antibody mediated restriction both *in vivo* and *in vitro* (Roberts *et al.* 1994; Bukreyev *et al.* 2008)(Bukreyev, Yang and Collins 2012). Neutralisation studies with monoclonal antibodies suggest that the F protein is the main target of cross-reactive neutralising antibodies (Olmsted *et al.* 1986). These studies show that only F specific monoclonal antibodies mediate complete neutralisation of RSV *in vitro* while antibodies to the G

protein mediate incomplete or strain specific neutralisation (Anderson, Bingham and Hierholzer 1988; Garcia-barreno *et al.* 1989). Multiple antigenic sites (I, II, IV and 0) in the F protein have been identified as targets of neutralising antibodies (López *et al.* 1990, 1998; McLellan *et al.* 2013). Other than the F and the G proteins, studies on the role of other RSV proteins in inducing resistance to RSV challenge have shown that while the M, P and SH did not induce detectable resistance to RSV challenge of BALB/c mice, the M2 and N induced significant but not complete resistance. (Connors *et al.* 1991). In addition to humoral responses, animal studies have reported the development of cytotoxic T-cell responses directed at the N, SH, F, M, M2 and NS2 proteins (Cherrie *et al.* 1992).

Several RSV proteins also inhibit innate immune responses. The 1 and NS2 proteins prevent induction of and disruption of IFN α/β activity (Spann *et al.* 2004). Studies with bovine and human RSV strains with a deletion of the SH gene have demonstrated the roles of the SH protein in the inhibition of TNF-α signalling, increased IL-1β response, and induction of apoptosis (Fuentes *et al.* 2007; Taylor *et al.* 2014; Russell *et al.* 2015). Cytokines are an important element of the early immune response to RSV. RSV infection elicits production of an array of cytokines that mediate a number of functions that are not only necessary for virus clearance but that may also promote pathology. T cells produce pro-inflammatory mediators in response to RSV infection. Type 1 T helper (Th1) cells produce IL-2, IFN-g and lymphotoxin while Th2 cells produce IL-4, -5, -6 and -10. The type of T helper response elicited in response to infection is largely dependent on the cytokine milieu present at the time of priming (Openshaw 2002). An imbalance in Th1/Th2 responses to RSV has been cited as a contributor to severe illness (Folkerts *et al.* 1998). RSV

associated disease severity appears to be the product of a Th2 skewed response (Becker 2006) although this has not been consistently confirmed in the respiratory secretions of infants with acute RSV bronchiolitis (Garofalo *et al.* 2001).

To a large extent, the dominant RSV strain in a particular epidemic is one that had been observed at low frequencies in previous seasons and has undergone some genetic changes since it was last seen (Cane and Pringle 1995; Yamaguchi *et al.* 2011) However, studies in Kilifi and other places have shown that re-infection with the same RSV strain (same G-gene sequence) is possible and could cause mild to severe RSV disease in infants (Hall *et al.* 1991; Venter, Collinson and Schoub 2002; Scott *et al.* 2006). In addition, a study of RSV isolates from Cuba in 1994-1995 showed that the isolates had only five G gene nucleotide differences with the RSV long strain isolated in 1956 in USA (Valdés *et al.* 1998). The lack of variation in the G (and potentially F) protein of re-infecting strains fails to accord with their role as targets of protective immunity and implies changes in other regions of the genome may also contribute to immune escape. Considered together, the results of these studies suggest that changes in parts of the genome that do not encode for surface exposed proteins may still exert potent effects on both virus fitness and potentially alter the effectiveness of protective host responses. Consequently, vaccine development programmes are likely to benefit from full-length genome studies, in which variations in the entire virus genome and not only in surface expressed proteins are analyzed. This will result in better understanding the targets of host immune responses and the virus-specific mechanisms that allow for evasion from these responses.

1.5 G-gene Duplication Variants

RSV has a non-segmented genome and does not show the abrupt antigenic changes that occur, for example, in the influenza A viruses as a result of genome reassortment (Brown *et al.* 1998; Webby *et al.* 2000; Newman *et al.* 2008). Furthermore, recombination has not been observed or reported in natural infections with RSV to date and hence might not be a source of diversity for the virus population. As with any RNA virus, the low polymerase fidelity in virus replication leads to high mutation rates and existence of virus quasi-species (Holland *et al.* 1982). The distribution of fitness amongst the swarm of a virus population both spatially and temporally has implications for its transmission. Rapid accumulation of sequence changes in the RSV G protein with time suggests selective forces acting on the viral population (Cane and Pringle 1995). However, antigenic variation is not necessarily the result of immune selection (they may be random) as neutral variation may also result in amino acid replacements over time (Domingo *et al.* 1993).

Two large duplication variants within the G-gene have been detected in RSV isolates, one in each group, *Figure 1.7*. The initial variant involved a 60-nucleotide duplication in RSV B that was first detected in Buenos Aires, Argentina, in 1999 and named the 'BA genotype' (Trento *et al.* 2003). Retrospective analysis of RSV samples from Madrid, Spain, detected the BA genotype between 1998 to 1999 (Trento *et al.* 2010). The BA genotype has spread globally and become the predominant group B genotype, and also undergone genetic drift resulting in further subdivision into genotypes BA1-BA10 (Trento *et al.* 2006; Dapat *et al.* 2010; van Niekerk and Venter 2011). The estimated nucleotide substitution rates for the BA viruses identified in Argentina were significantly higher than those previously reported for both RSV-A and B (Trento *et*

al. 2006). Notably, additional positively selected and glycosylation sites were identified within the duplicated segment in subsequent analysis (Zlateva *et al.* 2005).

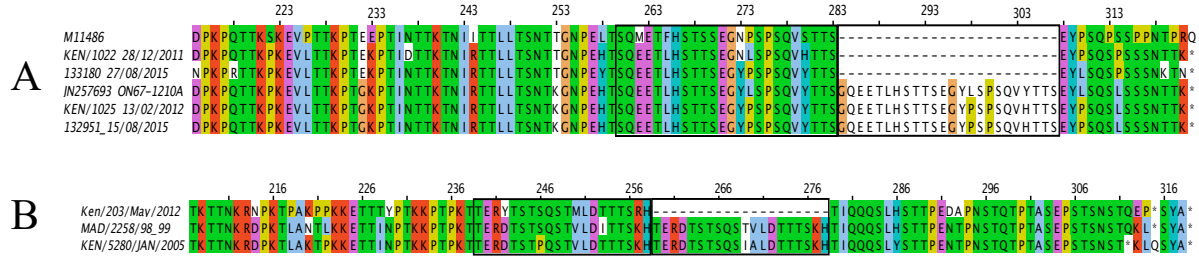


Figure 1.7: Amino acid alignments showing the duplication regions in the ON1 and BA genotypes.
The areas enclosed by black rectangles show the duplication regions within the G protein for genotype (*A*) ON1 and (*B*) BA viruses. The amino acid positions for both (*A*) and (*B*) are relative to references JN257693 and JF704220, respectively.

The more recent duplication variant involves a 72-nucleotide duplication within the G-gene of RSV-A. It was first identified in Ontario, Canada, in 2010 and named the 'ON1 genotype' (Eshaghi *et al.* 2012). It has subsequently been detected globally in several countries (Lee *et al.* 2012; Prifert *et al.* 2013; Tsukagoshi *et al.* 2013; Valley-Omar *et al.* 2013; Choudhary *et al.* 2013; Khor *et al.* 2013; Agoti *et al.* 2014b; Panayiotou *et al.* 2014; Pierangeli, Trotta and Scagnolari 2014; Ren *et al.* 2014; Auksornkitti *et al.* 2014; Avadhanula *et al.* 2015; Ahmed *et al.* 2016; Yoshihara *et al.* 2016; Fall *et al.* 2016; Korsun *et al.* 2017; Park *et al.* 2017; Calderón *et al.* 2017; Comas-García *et al.* 2018; Gaymard *et al.* 2018). Through continuous hospital surveillance for RSV in paediatric pneumonia admissions, the first detection of the ON1 genotype in Kilifi was in February 2012, and by the end of 2012 the ON1 variant was the dominant genotype of the RSV-A strains (Agoti *et al.* 2014b). However, it is neither known when this genotype was first introduced in Kenya nor all the variants circulating.

Most important for this thesis project, these duplications provide a vital tag as a mechanism for tracking a virus (in this case RSV) at a range of scales (local community, across the country, intra-continent and global) with the benefit of knowing how a new variant emerged, when it first entered a community, and assessing the pace and nature of spread and the accompanying genomic changes. From recent reports, it is becoming apparent that these G-gene duplication variants are not unique to RSV. Two duplications, 180nt and 111nt, have been detected in the G-gene of human metapneumovirus (hMPV) with increase in their frequencies in subsequent epidemics (Piñana *et al.* 2017; Saikusa *et al.* 2017a, 2017b). Therefore, further studies can be designed to not only determine the mechanisms leading to the generation of these duplications, but also the functional importance underlying the fitness and survival of these duplication variants.

1.6 RSV Whole-genome Studies

Previous and most of the current RSV epidemiological studies are based on partial G-gene sequencing. This is because the G protein interacts with the host cell receptors (thus far undefined), is a target for neutralizing antibodies, and is highly variable (Johnson *et al.* 1987). G protein variability is greater than that of other RSV proteins, both between and within the RSV-A and B groups (Sullender 2000). However, the portion of the G-gene used in the studies mentioned above, the C-terminal hypervariable region, accounts for only ~2% (300 nt) of the RSV genome.

There are few RSV studies at the whole genome level (Kumaria *et al.* 2011; Rebuffo-Scheer *et al.* 2011; Tan *et al.* 2012, 2013; Agoti *et al.* 2015a, 2017; Bose *et al.* 2015;

Schobel *et al.* 2016). Because Sanger chain-termination sequencing is costly and labour-intensive, earlier RSV whole genome sequencing (WGS) studies were limited to reference and mutant strains as part of functional biology and vaccine studies (Connors *et al.* 1995; Tolley *et al.* 1996). However, with the development and increasing access to next generation sequencing (NGS) technologies (Metzker 2010), in addition to development of optimized protocols for whole genome sequencing (Agoti *et al.* 2015a; Goya *et al.* 2018), it has become more affordable and efficient to sequence full genomes (Morey *et al.* 2013).

Very few studies have looked at diversity in genes other than the G (Johnson and Collins 1988, 1990; Moudy, Sullender and Wertz 2004; Agoti *et al.* 2015a). Since 2011, there has been interest by different groups working with RSV to sequence full genomes. These whole genome studies have not only provided novel insights on the basic biology of the virus but also elucidated the patterns of diversity in the intergenic regions, gene-start and gene-end sequences, where differences within these areas have been ascribed some functional importance especially in transcription and selection advantage (Kuo, Fearns and Collins 1997; Sutherland, Collins and Peeples 2001; Harmon and Wertz 2002; Kumaria *et al.* 2011; Rebuffo-Scheer *et al.* 2011).

The rate of nucleotide substitution for the G gene encoding the attachment protein has been estimated to be 1.83×10^{-3} (95% confidence interval [CI], 1.44 - 2.26×10^{-3}) and 1.95×10^{-3} (95% CI, 1.15 - 2.34×10^{-3}) nucleotide substitutions/site/year for group A and B, respectively, with some variation dependent on the timescale of observation (Zlateva *et al.* 2004, 2005). Similarly, although at a lower rate, there is also significant accumulation of substitutions across the rest of the genome (Agoti *et*

al. 2015a, 2017). At present, there is limited knowledge about the selective forces acting on genes other than the G gene as a result of paucity of WGS, particularly from the same location over a period spanning multiple seasons (Tan *et al.* 2012, 2013). Therefore, genetic signatures across the rest of the genome that might additionally inform on the adaptive mechanisms of RSV following introduction into communities have not been investigated before.

NGS platforms are now increasingly in use in clinical microbiology laboratories (Deurenberg *et al.* 2017). However, there are currently relatively few RSV genomes (~ 1,000) to obtain a clear understanding of the virus. A whole genome sequence of a virus obtained from one isolate at a particular point in time provides a complete snapshot of that virus at that time. Sequencing multiple genomes sampled at different time points could further inform on the genetic diversity of the virus, its evolutionary history, genes and sites under immune selection, and the fine-grain resolution of transmission patterns into, within and between populations. Human influenza virus studies, for example, for which there are over 3,000 genomes have reported epistatic interactions between the gene coding for the virus surface haemagglutinin (HA) protein with genes encoding other (non-surface) viral proteins (Mitnaul *et al.* 2000). Studying RSV genetic variation at the genome level could have the potential of revealing additional genes, other than the G-gene, and individual codons that are most crucial for RSV survival.

Tracking viral transmission over short periods and smaller geographical regions requires a greater evolutionary signal that is provided by whole genome sequences (Kamoun *et al.* 2015). Previous analyses have shown that the G-gene suffices to show

general patterns of diversity of RSV over multiple epidemics both locally and globally (Peret *et al.* 2000; Botosso *et al.* 2009; Katzov-Eckert *et al.* 2012). However, over shorter periods G-gene falls inadequate as the analyses include recently diverged isolates for which greater resolving power is required. Through our previous analyses, we observed multiple sequences sourced from the same epidemics that are 100% identical in the G-gene but possessing differences elsewhere in the genome (Agoti *et al.* 2015a). Such differences in the other open reading frames (ORFs) may be exploited to characterize transmission chains over short periods (single epidemics) or distances. In addition, as a result of the rapid spread of the virus, global variation equilibrium is quickly attained making tracking of transmission at the global level equally difficult over short periods using G-gene alone.

1.7 Evolution of genotypes and variants

RSV is continuously evolving as is exemplified by the rise of new variants and genotypes (Trento *et al.* 2010; Katzov-Eckert *et al.* 2012). The G-protein shows capacity to accommodate frequent multiple nucleotide and amino acid mutations and, twice in the recent past, large duplications. Factors such as short-term variant-specific herd immunity, selection, transmission bottlenecks and neutral epidemiological dynamics are likely to affect RSV genotype/variant prevalence (Sullender 2000; Botosso *et al.* 2009). With the complex RSV circulation patterns, little is known about when most of the genetic diversity observed is generated, for example between or within the annual epidemics.

Although minor duplications or deletions of 1-2 codons in the G protein of RSV have been reported prior to emergence of the 60 and 72 nucleotide duplication genotypes, they were identified in few epidemics and were short-lived (García *et al.* 1994;

Melero *et al.* 1997; Moura *et al.* 2004; Blanc *et al.* 2005; Trento *et al.* 2006). The BA genotype has been, and still is, the dominant RSV B genotype for more than 15 years globally while ON1 is spreading globally fast (Trento *et al.* 2010; Duvvuri *et al.* 2015). There are examples in several virus families (e.g. Orthomyxoviridae and Arenaviridae) of minor changes in viral genome sequences leading to large impacts on viral pathogenesis (Hatta *et al.* 2001; Conenello *et al.* 2007; Sullivan *et al.* 2011). Differential pathogenesis and disease severity arising from different RSV groups and genotypes is inconclusive (Walsh *et al.* 1997; Martinello *et al.* 2002; Stokes *et al.* 2011; Tran *et al.* 2013; Panayiotou *et al.* 2014; Otieno *et al.* 2017). The determinants of differential transmission and pathogenesis associated with particular RSV strains have been investigated and some evidence identified, e.g. better binding avidity in BA versus non-BA viruses (White *et al.* 2005, 2007; Villenave *et al.* 2012; Stokes *et al.* 2013; Hotard *et al.* 2015). However, it is important to note that the study by Hotard *et al* compared virus binding and replication of the BA genotype in cells expressing heparin sulphate compared to cells that don't while RSV uses CX3CR1 as a receptor to infect human ciliated airway epithelial cells (Johnson *et al.* 2015). Nonetheless, it is somewhat likely that the circulation and dominance of the RSV variants with large duplications over multiple epidemics points to likely pathogenesis and/or transmission fitness advantage. It is also possible that pre-existing herd immunity to previous variants in circulation gave a fitness advantage to the duplication variants to which there was less cross-protective immunity (immune selection) as suggested by a high dN/dS ratio within the immunogenic regions of RSV G (Botosso *et al.* 2009).

1.8 Viral phylodynamics; the basics

The success or failure of a pathogen is entirely dependent on its ability to survive and reproduce in one host, and spread to a new host or environment (Bliven and Maurelli

2016). Host immune systems, predators, microbial competitors, parasites, and environmental resource limitations all exert selective pressures that shape the genomes of microbial populations (Toft and Andersson 2010). Grenfell *et al.* coined the term "phylodynamics" to refer to this melding of immunodynamics, epidemiology, and evolutionary biology (Grenfell *et al.* 2004), *Figure 1.8*. Pathogen evolution is often characterized by accumulation of genetic variation, and this genetic variation is modulated by host immunity, transmission bottlenecks, and epidemic dynamics. It is within this phylodynamic framework that the current study aimed to infer the local and global molecular epidemiological dynamics of RSV.

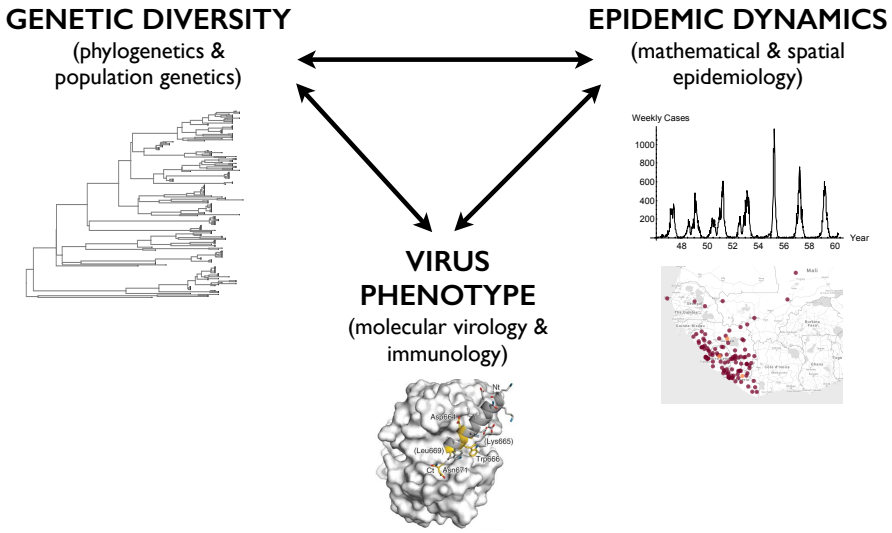


Figure 1.8: A representation of the different components of viral phylodynamics (Pybus 2016)

Phylodynamic data typically comprises gene or genome sequences sampled at different points in time and from different locations. The time points can be weekly or monthly within a single epidemic or over several epidemics (years). Locations can be subjects' homes, local hospitals, cities, countries, or even continents. Because of the commensurate time-scale of evolutionary and spatial dispersal dynamics, genetic data when combined with other data streams may offer a valuable source of information to

reconstruct transmission for rapidly evolving pathogens such as the emergent RSV ON1 genotype (Holmes 2008; Pybus and Rambaut 2009). The increasingly complex quantitative phylodynamic approaches allow integration with these other data types, such as phenotypic trait data (e.g. immunological assays) in these analyses (Wallace *et al.* 2007; Wallace and Fitch 2008; Lemey *et al.* 2009).

1.9 Phylogeography of respiratory viruses and determinants of spread

The worldwide population is increasingly urbanized and mobile fuelled by technological advances and ease of movement. When this is considered with respect to infectious pathogens (including RSV), it presents a heightened risk of infectious disease spread of global magnitudes (Brockmann 2010). In fact, a recent study reported detection of multiple respiratory pathogens on frequently touched surfaces at airports (often inevitable for embarking passengers for security purposes), potentially creating pathogen "pick-up" hubs for onward transmission (Ikonen *et al.* 2018). To limit the enormous economic and social impacts from potential infectious disease global pandemics, for example by proposing appropriate containment and disease spread prevention strategies, a thorough understanding of the dynamics of spatial spread is critical. While RSV does not cause pandemic-sized outbreaks, it's recurrence and persistence within communities poses a significant global health burden (Shi *et al.* 2017).

A study of emerging infectious diseases (EIDs) between 1940 and 2004 by Jones *et al*. reported that the origins of the EIDs were significantly correlated with socio-economic, environmental and ecological factors (Jones *et al.* 2008). This study, together with many others (Alonso *et al.* 2007; Lam *et al.* 2013; Obando-Pacheco *et*

al. 2018), make a case for more research in order to illuminate on where new EIDs are most likely to originate (emerging disease 'hotspots') as well as the source of current and recurrent human infections. Further, these studies may help global public health policy makers to allocate resources more appropriately to regions where EIDs are most likely to originate for control purposes. Obando-Pacheco *et al.* recently used national RSV surveillance reports and medical databases from 27 countries to describe the timings of RSV epidemics worldwide and concluded that the global annual RSV epidemics started in the South hemisphere moving to the North (Obando-Pacheco *et al.* 2018). However, the authors admit that RSV surveillance data was scarce particularly from middle- and low-income countries that lack RSV surveillance networks, yet this is where most RSV burden rests, and therefore more information is critical to obtaining a better picture of global RSV epidemic patterns.

Phylogeography or phylogeographic methods enable inference of the geographical history of genetic lineages, and therefore can shed light on the transmission dynamics of a given pathogen. Two stochastic models of phylogenetic diffusion have been proposed to perform spatiotemporal reconstructions in a Bayesian framework (Bloomquist, Lemey and Suchard 2010; Faria *et al.* 2011): A continuous time Markov chain (CTMC) process to model transitioning among discrete location states throughout evolutionary history (Lemey *et al.* 2009) and a Brownian random walk process to model diffusion in continuous space (Lemey *et al.* 2010). The discrete model has been extended beyond estimating the migration history of a virus of interest to testing and quantifying a range of potential predictors of spatial spread (Lemey *et al.* 2014). To the best of our knowledge, no study has reported a source-sink type transmission model for RSV or rather an in-depth phylogeographic analysis of global and local (sub-country and country level) RSV spread, and the potential

predictors of such spread. Such a study would significantly build on the surveillance report-based studies such as that by Obando-Pacheco *et al.* as far more countries deposit RSV time-stamped sequence data without epidemiological details of RSV epidemics.

1.10 Surveillance of respiratory viruses in Kenya

Surveillance of different respiratory viruses is ongoing in several parts of Kenya (Adazu *et al.* 2005; Odhiambo *et al.* 2012; Scott *et al.* 2012). The major institutions in these surveillance activities are the Kenya Medical Research Institute (KEMRI), Ministry of Health, Kenya Ministry of State for Defence (KMoD), KEMRI-Wellcome Trust Research Programme (KWTRP), Centres for Disease Control and Prevention - Kenya (CDC-K), and US Army Medical Research Directorate – Kenya (USAMRU-K). The Virus Epidemiology and Control (VEC) group has established collaborative links with CDC-K previously through a study on the RSV genetic diversity at the Dadaab refugee camp (Agoti *et al.* 2014a) and currently through a countrywide pathways of transmission study titled SPReD (Studies of the Pathways of transmission of Respiratory virus Disease) (http://virec-group.org/spred-kenya/). The collaboration enables use of CDC-K archived and RSV positive samples collected from different parts of the country for epidemiological and molecular evolutionary studies.

1.11 Justification/contribution of the proposed study to Knowledge

RSV is an important cause of childhood acute severe lower respiratory tract illnesses (ALRTI) and a major contributor to hospital admission of infants. The availability of a vaccine would significantly lower the disease burden that is most experienced in

developing countries, yet none exist at present. RSV circulating in a community appears to change season by season and it is thought that this ability of RSV to cause repeat infections and recurrent epidemics is driven by genetic and antigenic variation within the virus. An opportunity has arisen to best characterize the dynamics of RSV transmission and molecular evolution at various scales (local community, countrywide, continental and global) due to the coincidence of three factors, namely; the occurrence of a new RSV variant with a trackable mutation tag (the ON1 genotype), the availability of respiratory virus samples across Kenya from a period dating back to the first introduction of the genotype, and advancement of sequencing technologies enabling complete virus sequence characterization. In addition, the data generated is potentially useful for formulating and predicting the impact of RSV disease intervention strategies.

1.12 Hypothesis

Using partial G gene and whole genome sequence data of the RSV ON1 genotype following first introduction to Kenya will elucidate the key characteristics associated with RSV introduction, spread and persistence within a population.

1.13 Study objective

The overall objective is to use the recently emerged genotype ON1 as a unique tag to characterize the diversity, evolution and phylogeography of RSV in Kenya and thereby develop improved understanding of the nature and factors important to virus spread and persistence.

1.13.1 ***Specific objectives***

1. Study the nature and pace of genomic variation of genotype ON1 variants in Kenya.

   The aim was to identify sequence signatures that characterize the ON1 variants in comparison to prevailing group A genotypes, how such substitutions may relate to fitness advantage, and compare the rate of evolution for the ON1 genotype to those previously estimated for other group A and B genotypes in Kilifi and globally.

2. Determine the spatiotemporal dynamics of the spread of the ON1 variants.

   The goal of this objective was to unravel local RSV epidemic seeding patterns, e.g. whether only a single variant or multiple variant introductions may be required to cause a local RSV epidemic, and also if the persistence of the virus in a community is fuelled by local diversification of introduced variants or frequent introductions from outside the community.

3. Understand the connectivity of RSV epidemics at various levels.

   This objective aimed to discern RSV transmission patterns e.g. whether different locations experience independent virus introductions or spread from one location to another. Ultimately, this would highlight on the dispersal rate and presence of any identifiable pathways of spread of RSV both locally and globally.

4. Identify factors influencing RSV transmission dynamics.

The target of this objective was to deduce the socio-ecological factors (e.g. geographic distance, travel, population size and density, etc), other than genetic factors, that could potentially influence RSV transmission patterns at different scales.

1.14 Manuscripts the candidate contributed during his PhD study period

Here, I list papers that have either been published or in preparation and are part of the PhD project, as well as papers that I contributed to while undertaking the PhD and relevant to my field of research.

1.14.1 ***Published and part of the thesis***

Otieno JR, Kamau EM, Oketch JW, Ngoi JM, Gichuki AM, Binter Š, Otieno GP, Ngama M, Agoti CN, Cane PA, Kellam P, Cotten M, Lemey P, Nokes DJ (2018) Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evol. 2018 Sep 24;4(2):vey027. doi: 10.1093/ve/vey027*

Otieno JR, Kamau EM, Agoti CN, Lewa C, Otieno G, Bett A, Ngama M, Cane PA, Nokes DJ (2017) Spread and Evolution of Respiratory Syncytial Virus A Genotype ON1, Coastal Kenya, 2010–2015. *Emerg Infect Dis. 2017 Feb;23(2):264-271. doi: 10.3201/eid2302.161149*

Otieno, JR, Agoti CN, Gitahi CW, Bett A, Ngama M, Medley GF, Cane PA, Nokes DJ (2016) Molecular evolutionary dynamics of respiratory syncytial virus group A in recurrent epidemics in coastal Kenya. *J Virol. 2016 Apr 29;90(10):4990-5002. doi: 10.1128/JVI.03105-15. Print 2016 May 15.*

1.14.2 ***In preparation and part of the thesis***

Otieno JR, Gichuki AM, Lidechi S, Onyango C, Verani J, Poletto C, Agoti CN, Nokes DJ, Lemey P (2019) Local and Global Transmission dynamics of RSV.

1.14.3 ***Subsidiary but relevant and contributed to while doing the PhD***

Kiyuka PK, Agoti CN, Munywoki PK, Njeru R, Bett A, **Otieno JR**, Otieno GP, Kamau E, Clark TG, van der Hoek L, Kellam P, Nokes DJ, Cotten M (2018) Human Coronavirus NL63 Molecular Epidemiology and Evolutionary Patterns in Rural Coastal Kenya. *J Infect Dis. 2018 Mar 21. http://dx.doi.org/10.1093/infdis/jiy098*

Agoti CN, Munywoki PK, Phan MVT, **Otieno JR**, Kamau E, Bett A, Kombe I, Githinji G, Medley GF, Cane PA, Kellam P, Cotten M, Nokes DJ (2017) Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. *Virus Evol. 2017 Mar 11;3(1):vex006. doi: 10.1093/ve/vex006. eCollection 2017 Jan*

Agoti CN, **Otieno JR**, Ngama M, Mwihuri AG, Medley GF, Cane PA, Nokes DJ (2015) Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. *J Virol. 2015 Nov;89(22):11630-42. doi: 10.1128/JVI.01972-15. Epub 2015 Sep 9.*

CHAPTER TWO

2 Materials and Methods

2.1 Introduction

This chapter describes the study locations and populations from which samples were collected, and the laboratory and analysis methods used. However, for better readability, some methods are not described here but within the related chapters.

2.2 Study locations

For this thesis project, samples were collected through KWTRP and CDC-K from different sample collection sites across Kenya are shown in *Figure 2.1*.

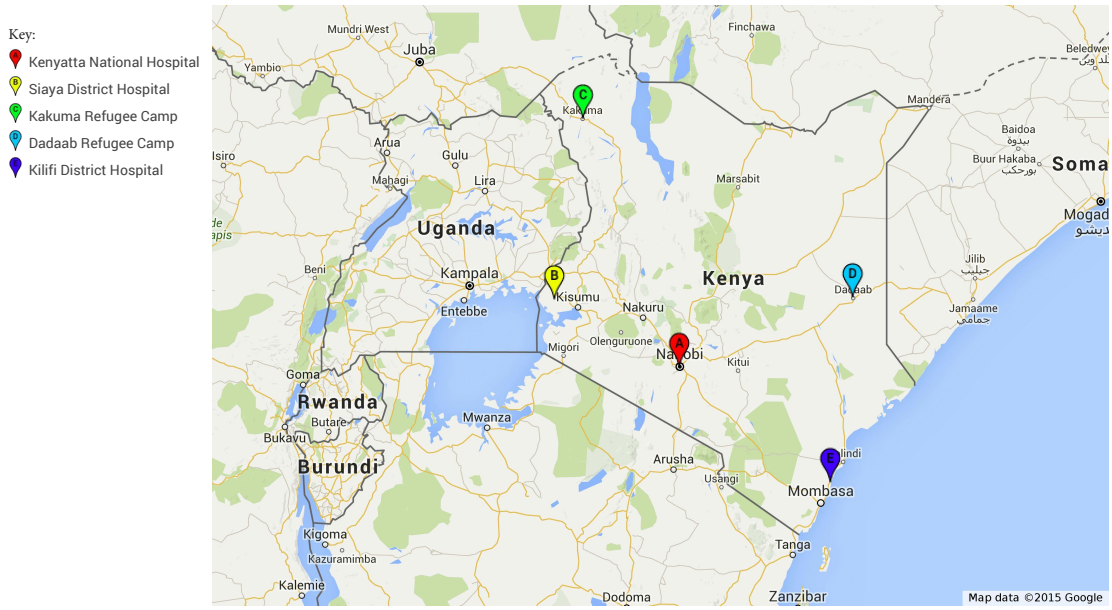


Figure 2.1: Respiratory viruses surveillance sites across Kenya from which RSV genotype ON1 samples were collected, 2011 - 2016

The Sites are marked A to E, with the key shown top left.

2.2.1 ***Kilifi County***

Kilifi County is located at the Kenyan coast with a rural (predominant) and semi-urban population of approximately 1.1 million (Kenya National Bureau of Statistics 2013). The area experiences a tropical climate with two rainy seasons (long rains from May to July and short rains in October and November), with the main economic activities being subsistence farming (maize, cassava, cashew nuts, coconuts, goats and dairy cattle) and fishing. The County has a child rich population, where 0-14 year olds constitute 47% of the total population (Kenya National Bureau of Statistics (KNBS) and Society for International Development – East Africa (SID) 2013).

The majority of the epidemiological samples described herein were collected within the Kilifi Health and Demographic Surveillance System (KHDSS) (Scott *et al.* 2012) with a minor proportion from outside the KHDSS but within the County. The KHDSS area was defined and mapped for demographic surveillance, clinical and epidemiological research by KWTRP in the year 2000. It covers an area of 891 Km^2, 50 km north and south, and 30 km west of the main referral hospital within the County, the Kilifi County Hospital (KCH). In addition to the KCH, there are 20 public health facilities operated by Kenya's Ministry of Health offering outpatient services within the KHDSS. A population of around 296,000 residents (census 2016) are routinely monitored through household enumeration visits conducted every 4 months. It has been previously estimated that 60% of the infant and young children admissions to KCH comes from the KHDSS area (Moïsi *et al.* 2011).

The Virus Epidemiology and Control (VEC) group at KWTRP runs an ongoing hospital-based RSV surveillance since 2002 at KCH (Nokes *et al.* 2009). Surveillance

of respiratory viruses has also been conducted in households and additional Health Centres within Kilifi County (Nokes *et al.* 2008; Munywoki *et al.* 2011, 2014; Nyiro *et al.* 2018). The active surveillance is supported by the KHDSS that links medical data obtained from health facilities to socio-demographic and geo-positional data obtained through the triannual KHDSS enumerations and decennial national population and housing census (Scott *et al.* 2012). This active surveillance has enabled quantification of respiratory virus disease burden in Kilifi County, characterization of virus transmission within households and schools and also a description of the molecular evolution of these viruses within Kilifi County.

2.2.2 ***Other regions of Kenya (CDC-K surveillance sites)***

The CDC-K conducts surveillance for influenza and influenza-like-illnesses (ILI) and Severe Acute Respiratory infections (SARI), e.g. RSV, adenovirus, human metapneumovirus, parainfluenza and enteroviruses through several clinical surveillance sites throughout Kenya (Bigogo *et al.* 2013; Emukule *et al.* 2014; Katz *et al.* 2014). These include sentinel regional hospitals, health facilities at demographic surveillance sites and refugee camps with varied demographic characteristics; urban, rural, high mobility, low socio-economic communities, etc (Adazu *et al.* 2005; Feikin *et al.* 2011; Odhiambo *et al.* 2012). The sentinel hospital surveillance sites were set up by KEMRI-CDC-Kenya and Ministry of Health as part of Global Influenza Program and have been in operation since 2006. However, the RSV studies by CDC-K primarily target disease burden with no analysis characterizing RSV molecular evolution and phylogeography in the various sampling sites.

2.3 Clinical specimens

Clinical data related to the specimen to be processed was extracted from existing data capture systems in accordance with the respective study site. In both the CDC-K sentinel surveillance hospitals and KCH either or both nasopharyngeal (NP) and oropharyngeal (OP) swabs are collected.

2.4 The study designs and population

The samples analyzed as part of this PhD project were from the three studies described hereafter.

2.4.1 ***RSV inpatient (IP) study (2011-2016)***

The RSV IP study is a long-term surveillance study of respiratory viruses within Kilifi County, starting 1st January 2002 to present (Nokes *et al.* 2009). While the initial objectives of the study were specific to RSV and intended to quantify the burden of disease requiring hospitalization, define the epidemiological patterns, determine social contact patterns, and support immunological and molecular epidemiological investigations in this developing country setting (Scott *et al.* 2004; Nokes *et al.* 2008, 2009; Sande *et al.* 2013; Kiti *et al.* 2014; Nyiro *et al.* 2017), additional human respiratory viruses such as metapneumovirus (HMPV) (Owor *et al.* 2016), rhinovirus (HRV) (Onyango *et al.* 2012b), influenza (Onyango *et al.* 2012a), and coronavirus (HCoV) (Kiyuka *et al.* 2018) were subsequently included. Samples are collected from children (under 5 years of age) admitted to KCH presenting with syndromically defined severe or very severe pneumonia according to the World Health Organization (WHO) criteria (Nokes *et al.* 2009).

The samples analyzed in this thesis project from the RSV IP study were collected between September 2011 and August 2016, *Figure 2.2*. During this time period, a total of 3,157 samples were collected from eligible children at KCH, 3,146 (99.7%) were tested for RSV by IFAT and real-time PCR, and 801 (25.5%) RSV positives identified by either or both methods. Of these, 54.2% (423/740) were RSV-A, 39.6% (305/740) RSV-B, and 1.6% (12) RSV-A/B co-infections by real-time PCR.

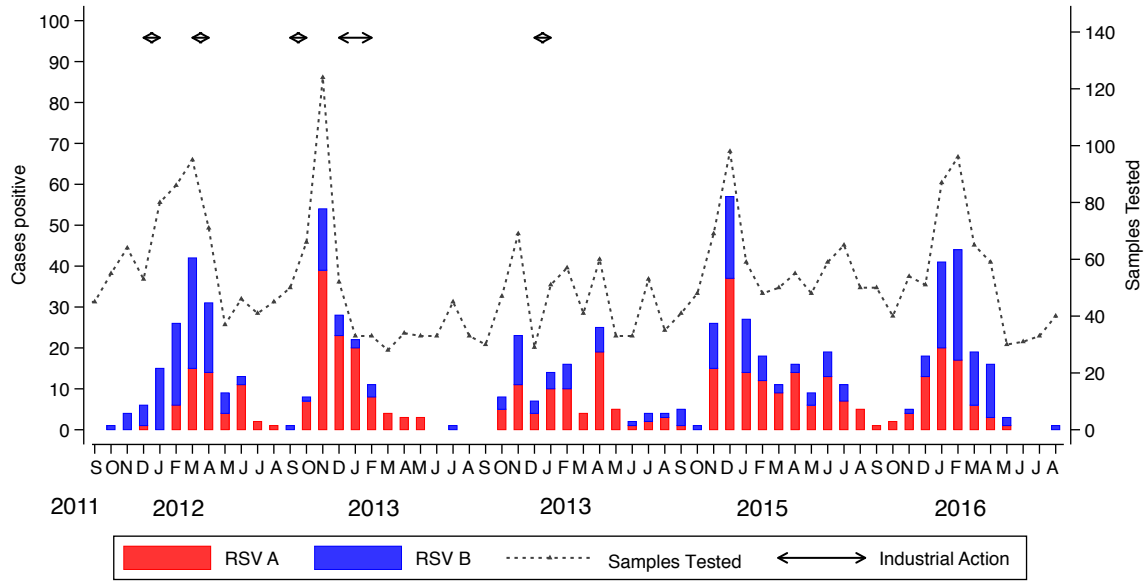


Figure 2.2: Temporal patterns of RSV strains from Kilifi Sept. 2011 – Aug. 2016.
The combined monthly detection frequency of RSV-A (red) and RSV-B (blue) viruses from the KCH child in-patient surveillance study, September 2011 to August 2016. The black dashed line shows the number of samples tested while the solid black lines at the top show periods in which there were industrial action at KCH.

2.4.2 ***SPReD-Kenya Study (2011-2014)***

The SPReD-Kenya study (http://virec-group.org/spred-kenya/) is part of the larger SPReD study that aims to advance understanding of the nature of spread (i.e. characteristic routes of virus introduction, spread, persistence and fade-out) of respiratory viruses (including RSV, influenza, coronavirus and rhinovirus) at different scales of observation; from the individual, to the household and school, to the local community, to the country level, and across the continent. The information obtained from the study would also be used to innovate interventions. The work represents an integration of epidemiological, virus sequence, contact and mobility data.

This is a collaborative project between KWTRP and KEMRI/CDC-Kenya and aimed to collect and analyze approximately 7,000 nasal specimens per calendar year (2014-2016) from ten different sentinel surveillance sites across Kenya from patients of various ages with SARI or ILI. The ten sites were: Kenyatta National Hospital, Nyeri County and Referral Hospital, Siaya County and Referral Hospital, Lwak Mission Hospital, Kakuma Refugee Camp Clinics, Kakamega County and Referral Hospital, Mombasa County and Referral Hospital, Kibera (Tabitha outpatient) Clinic, Nakuru County and Referral Hospital, and Kilifi County Hospital, *Figure 2.3*.

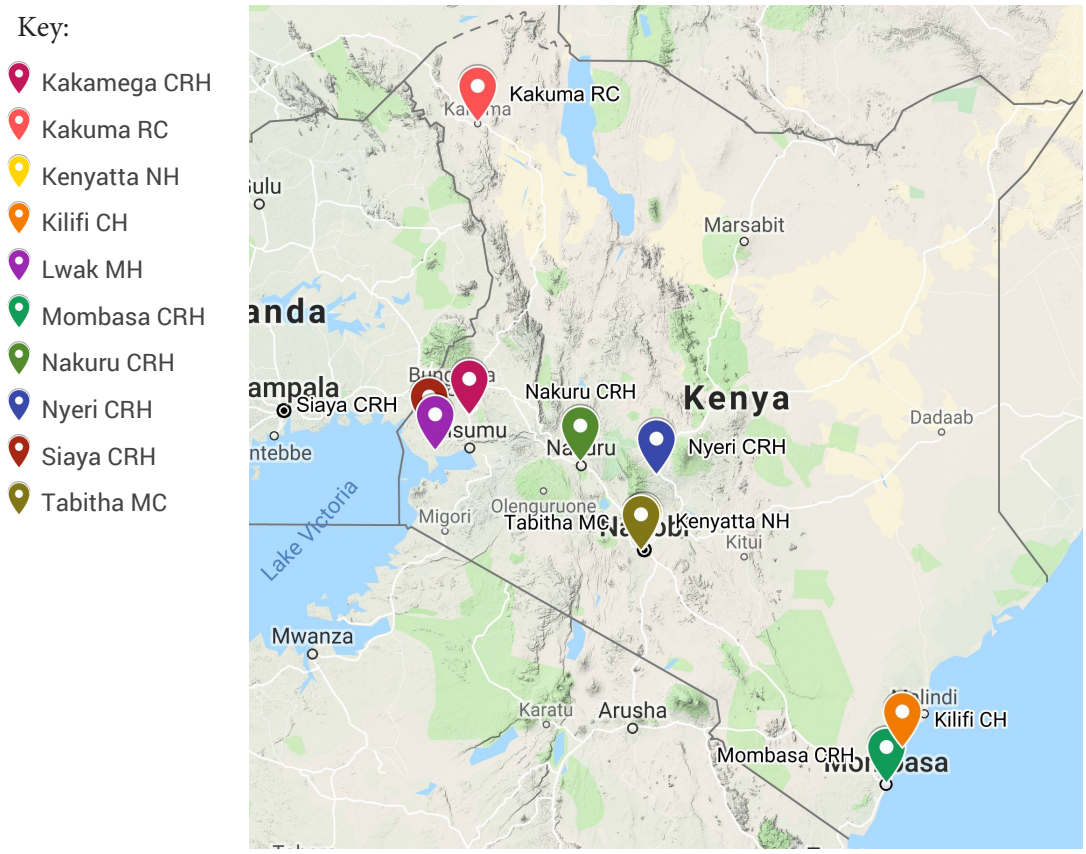


Figure 2.3: Map showing the SPReD-Kenya surveillance sites.

However, since it was of interest to determine when the ON1 viruses entered Kenya, and (i) the first detected case of ON1 in Kilifi was in February 2012 (Agoti *et al.* 2014b) while (ii) the first reported case of ON1 was from Ontario Canada in October 2010 (Eshaghi *et al.* 2012), the aim was to sample SARI and ILI specimens collected by CDC-K from January 2011 to December 2014 and targeted a minimum of one location each in Western, Central and Northern Kenya. The sites selected were Siaya (which included samples from Ting'wang'i and Lwak Mission Hospital), Kakuma Refugee Camp, Dadaab Refugee Camp, Kenyatta National Hospital and Kilifi, *Figure 2.1* and *Table 2.1*.

Table 2.1: Description of the selected ON1 study sites, patient inclusion criteria and sample types collected

#	Location	Setting	Inclusion criteria	Sample type(s)
1	Kenyatta National Hospital	Hospital Based	ILI[1] andSARI[2] in paediatric wards	NP/OP swabs
2	Siaya County Hospital	Hospital Based	SARI all patients*	NP/OP swabs
3	Dadaab Refugee Camp	Hospital Based	SARI all patients*	NP/OP swabs
4	Kakuma Refugee Camp	Hospital Based	SARI all patients*	NP/OP swabs
5	Kilifi County Hospital	Hospital Based	LRTI in under 13 years	NP/OP swabs

[1] ILI is defined as an acute respiratory illness with a measured fever of ≥38°C AND a cough with an onset of symptoms within the last 7 days.

[2] SARI is defined as an acute respiratory illness requiring hospitalization with a history of fever or measured fever ≥38°C AND a cough with an onset of symptoms within the last 14 days.

*In both adult and paediatric wards

NP - Nasopharyngeal; OP – Oropharyngeal;

*2.4.3 **SPReD-KHDSS study (2016)***

The SPReD-KHDSS study (http://virec-group.org/local-spred/) is also part of the larger SPReD study, with a specific focus on the KHDSS and similarly aiming to map patterns of spread of a range of respiratory viruses using epidemiological and nucleotide sequence data. In this study, samples were collected between January and December 2016 from patients of all ages presenting at nine purposively selected health facilities within the KHDSS with one or more acute respiratory infection (ARI) symptoms of cough, sneezing, nasal congestion, difficulty breathing, or increased respiratory rate for age as defined by the WHO (Nyiro *et al.* 2018). However, new-

borns aged less than 7 days and patients with ARI for more than 30 days were excluded. The nine public health facilities were (*Figure 2.4*); Matsangoni, Ngerenya, Mtondia, Sokoke, Mavueni, Jaribuni, Chasimba, Pingilikani and Junju, and their selection was based on representation across the geographical region, coverage of major road networks and variation in population density, (Nyiro *et al.* 2018).

Patient recruitment and specimen collection has been described in detail in Nyiro *et al* (Nyiro *et al.* 2018). Briefly, the recruitment of patients into the study and collection of specimens was integrated within the routine patient care at the nine selected outpatient facilities led by a resident clinician or nurse. Each facility had one or two sampling days per week between Monday and Friday (9.00 am - 1.00 pm). Qualifying patients were consented, and nasopharyngeal swab samples collected. A target of 15 samples per site per week was used on a 'first-come first-served' basis.

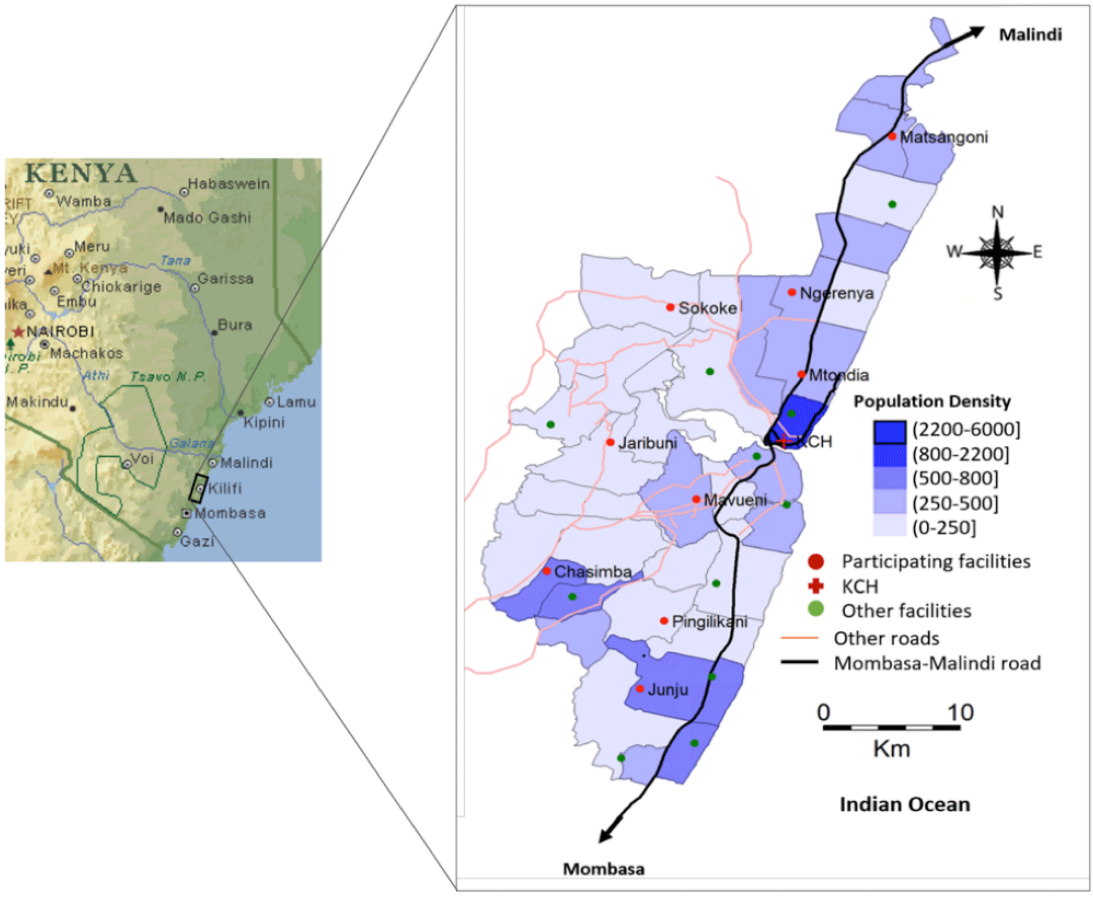


Figure 2.4: A map of the SPReD KHDSS study sites extracted from Nyiro *et al.* 2018
The map defining the KHDSS is expanded from the map of Kenya and shows population density (person per Km^2) and the nine health facilities (red circles) where the study was conducted in 2016. The green markers show the other public health facilities within the KHDSS area (Nyiro *et al.* 2018).

2.5 Sample size determination

For the period of 1st January to 31st December 2012, 834 nasopharyngeal swab specimens were collected from children eligible for the RSV IP study in Kilifi. Of the 834 samples, 240 (28.8%) were RSV positive: 123 (51.3%) were group A infections, 114 (47.5%) were group B infections, and 3 (1.3%) were A/B co-infections. Of the 126 combined group A and group A/B viruses, 104 (82.5%) were successfully sequenced in the G gene ectodomain region, and of these, 77 (74.0%) were of the ON1 genotype. Based on these observations, the expectation was that the majority of the group A viruses from 2013 onwards to be of genotype ON1. In total, the aim was to obtain full-length genomes from ≈200 samples from Kilifi.

As previously stated, a minimum of one location each in Western, Central and Northern Kenya was targeted for sampling. The *Table 2.2* shows the number of RSV positive samples available from the selected five sites for the period 2011-2014. For KNH 2011-2012 whereby respiratory samples were not screened for RSV, and approximately 700 samples usually collected each year, it was estimated that 30% of these samples would be RSV positive, i.e. 0.3(700X2) = 420 samples. This made an estimated the total of the RSV positives from the CDC-K sites to be 1,271, *Table 2.2*. Because RSV groups A and B are known to alternate in dominance, albeit with group A occurring more frequently, and using data collected from Kilifi over 12 years

(Agoti *et al.* 2015b; Otieno *et al.* 2016), it was anticipated that ~60% of the CDC-K RSV positive isolates would be RSV-A, i.e. 0.6(1271) = 763 specimens. Further, it was predicted that very few (~30%) of the RSV-A specimen from 2011 would be ON1 considering the genotype's initial detection in Kilifi in February 2012. Finally, it was expected that about two-thirds of the RSV-A positive specimens would have low Ct values (high viral load), e.g. Ct value of <30, which is suitable for WGS (Agoti *et al.* 2014a). Therefore, a total of ~400 isolates from western, central and northern parts of Kenya were anticipated to be taken through WGS.

Table 2.2: RSV positive specimens from 5 sites in Kenya conducting respiratory virus surveillance, 2011 - 2015

	2011	2012	2013	2014	**TOTAL**
Kenyatta National Hospital	TBS	TBS	12	90	**102**
Siaya County Hospital	125	101	80	52	**358**
Kakuma Refugee Camp	112	60	43	21	**236**
Kilifi County Hospital$^{\beta}$	N/A	126	66	119	**311**
Dadaab Refugee Camp	99	1	21	34	**155**
TOTAL	**336**	**288**	**222**	**316**	**1162**

TBS: To be screened for RSV. Approximately 700 samples collected per year.

$^{\beta}$Total number of RSV-A positive specimen from Kilifi.

N/A: Not applicable as no ON1 viruses were detected in Kilifi prior to 2012.

The data on the four CDC-K sites selected were clearly subject to variable collection effort over time and between sites. However, as a whole the sample set provided representation of RSV strains across the country in each of the years, and since the samples were already screened for RSV, they represented a cost-effective approach for this thesis investigation. Based on our previous studies, these numbers would be sufficient to identify seasonal patterns of RSV (Agoti *et al.* 2015a).

2.6 Study scientific and ethical approval

Samples used in this study that had previously been collected and stored had received scientific and ethical approval from the UK (Coventry Research Ethics Committee), US and Kenyan (KEMRI Scientific and Ethics Review Unit [SERU]) ethics committees: CDC-K/Ministry of Health Influenza Sentinel Surveillance Program (SSC No. 2558 and 2692) and KWTRP (SSC No. 1055 and 1433) (Nokes *et al.* 2008, 2009). The SPReD-Kenya study was approved by SERU (SERU No. 3044) while the SPReD KHDSS study was approved by both SERU (SERU No. 3103) and the University of Warwick Biomedical and Scientific Research Ethics Committee (BSREC# REGO-2015-6102) (Nyiro *et al.* 2018). An informed consent (written or verbal) depending on study site was obtained from the study patient or their guardian.

However, to use the samples above in this study further scientific and ethical approval was sought. KWTRP is located within a Centre of the Kenya Medical Research Institute known as the Centre for Geographic Medicine Research – Coast (CGMR-C). Initial approval for this study was sought at Centre level through the Centre Scientific Committee (CSC) that reviews all proposals for studies conducted at the Centre involving human subjects. Subsequently, upon approval by CSC, national scientific and ethical approval was sought from SERU for the committee meeting on 18/12/2015. An initial response letter from SERU was received on 12th January 2016, which suggested minor clarifications and changes. Following submission of the revised protocol, the study received scientific and ethical approval (SERU No. 3177) through a letter (*Appendix 7.1*) dated 4th February 2016.

2.7 Laboratory methods

2.7.1 ***RSV diagnosis***

All respiratory specimens analyzed in this project from the RSV IP study were diagnosed for RSV by antigen detection using indirect immunofluorescent antibody assay test (IFAT) (Light diagnostics, Chemicon, UK) and by nucleic acid quantitative real-time polymerase chain reaction (PCR) using either an in-house multiplex reverse transcription (RT) PCR (Gunson, Collins and Carman 2005) or the Fast Track Diagnostics Respiratory Pathogens 33 test [FTD Resp-33; Fast-track Diagnostics, Sliema, Malta] (Hammitt *et al.* 2012). The real-time RT-PCR further types the RSV positives by group, i.e. RSV-A and RSV-B. For the SPReD-Kenya and SPReD-KHDSS samples, only the real-time RT-PCR method was used for RSV detection (Nyiro *et al.* 2018). The 2011-2013 respiratory samples from CDC-K surveillance sites had previously been screened for RSV positivity using Taqman Array Cards but without RSV group information (Katz *et al.* 2012; Bigogo *et al.* 2013; Emukule *et al.* 2014).

2.7.2 ***RNA extraction***

For the RSV IP study samples, viral RNA was freshly extracted from 140µL of clinical specimens using QIAmp viral RNA mini kit (Qiagen Ltd, UK) in accordance with the manufacturer's protocol. In the final elution step, 60µL of viral RNA was collected. As for the SPReD-Kenya and SPReD-KHDSS samples, RNA was extracted from the respiratory specimens by Qiacube HT using an RNeasy extraction kit (Qiagen, Germany). Positive (RSV A2 strain culture supernatant) and negative (either a previous confirmed negative sample or RNase free PCR purity water) controls were included in every extracted batch of samples to monitor potential cross-contamination during sample processing and false negatives in case of process failure.

2.7.3 ***G gene RT-PCR and Sequencing***

Viral RNA reverse transcription and G gene PCR amplification was performed as previously undertaken and described by the group (Scott *et al.* 2004; Agoti *et al.* 2012, 2015b; Otieno *et al.* 2016). Briefly, extracted viral RNAs were reverse transcribed and amplified in a one-step reaction protocol (QIAGEN, Ltd) with the primers AG20 and F164 (*Table 2.3*) targeting the entire RSV G gene and part of the F gene. A microlitre of the resultant products was further amplified in a nested PCR with the primers BG10 and F1, *Table 2.3*. Success in amplification was confirmed on a 2% agarose gel (expected band size of ~ 830 bp) and sequencing done using the BigDye 3.1 Chemistry with the nested PCR primers and additional RSV-A group specific primers, 523F and 523R (*Table 2.3*). Sequenced contigs were assembled to obtain the consensus sequences using Sequencher v5.0.1 (Gene Codes Corporation, USA), Gap4 release 2.0.0b9 (Bonfield, Smith and Staden 1995) and/or Geneious v11.1.2 (Kearse *et al.* 2012).

Table 2.3: Primers for RSV-A G gene amplification and sequencing

Primer	Group	Gene	Position*	Function	Sequence (5'-3')
AG20	A/B	G	1-20	PCR	GGGGCAAATGCAAACATGTCC
F164	A/B	F	164-187	PCR	GTTATGACACTGGTATACCAACC
BG10	A/B	G	154-173	PCR & sequencing	GCAATGATAATCTCAACCTC
F1	A/B	F	3-22	PCR & sequencing	CAACTCCATTGTTATTTGCC
523F	A	G	538-557	Sequencing	ATATGCAGCAACAATCCAAC
523R	A	G	557-538	Sequencing	GTTGGATTGTTGCTGCATAT
533F	B	G	532-551	Sequencing	TGTAGTATATGTGGCAACAA
533R	B	G	532-551	Sequencing	TTGTTGCCACATATACTACA

*The positions of the primers are given relative to the A2 strain (M11486).

2.7.4 ***Development of the RSV whole genome sequencing method***

2.7.4.1 *The 6-amplicon WGS method*

We have previously undertaken whole genome sequencing of RSV within the VEC group in collaboration with and at the Sanger Institute (UK) through a novel 6-amplicon polymerase chain reaction (PCR) amplification strategy, *Figure 2.5A*, followed by amplicon sequencing using Illumina MiSeq (Agoti *et al.* 2015a). The primers had lengths of between 20-28 bases and a calculated melting temperature (Tm) of 56.9-58.4°C, *Appendix 7.2*. However, sequencing difficulties were experienced in the 3' and 5' ends of the genomes. In addition, since ON1 was a novel genotype, the previously designed group A primers were mapped onto available genotype ON1 genomes and some primer mismatches both at the ends and internally within the genomes were observed. The WGS method was transferred to Kilifi and performed an update of the existing primers with the mismatches using the same bioinformatics approach previously described (Agoti *et al.* 2015a).

2.7.4.1 *The 14-amplicon WGS method*

With help from Matthew Cotten (previously at the Wellcome Sanger Institute), we developed a 14-amplicon WGS method (*Figure 2.5B*). All Human orthopneumovirus sequence entries with lengths in the range of 2000-16000 nt were retrieved from GenBank (On 5th Dec 2015). Patent entries and entries with multiple Ns were excluded. In addition, RSV genome sequences from Kilifi studies not yet in GenBank at the time (5th Dec 2015) were included to generate a final set of 717 entries with a total sequence length of 9568447 nt. All possible 23 nt sequences were generated and then trimmed to a final calculated Tm of 47.9-49.5 °C. Sequences with homology to rRNA sequences, with GC content outside of the range of 0.3 to 0.75 or with a single

nucleotide fractional content of >0.6 were removed. The primer set was made non-redundant to yield a set of 98001 potential primer sequences (PoPros). The PoPros were mapped against all available RSV-A genomes, the number of perfect matches (frequency score) was retained as a score of sequence conservation. Finally, the RSV genome sequence was divided into 14 overlapping amplicons with 217 nt overlaps spanning the virus genome. The terminal region of each amplicon was defined as a primer bin. PoPros that mapped within each primer bin were identified and the two PoPros with the highest frequency score were selected. PoPros that mapped to the reverse bins were reverse complimented. In this manner, a final set of 58 primers was prepared. The primer sequences, lengths, calculated melting temperature), fractional GC content and mapping position on the RSV genome are presented in *Appendix 7.3*.

2.7.4.1 *Reverse transcription and PCR amplification*

Reverse transcription of RNA molecules and PCR amplification were performed initially with a 6-amplicon six-reaction strategy (Agoti *et al.* 2015a) and later using either a 6 or 14-amplicons strategy split into two reactions of three [1,3.5 and 2,4,6] and seven [1,3,5,7,9,11,13 and 2,4,6,8,10,12,14] amplicons, respectively. The reduction in the number of reactions was to cut down on the PCR amplification costs as amplification success was similar to the previous six reactions. Splitting the two reactions into odd and even numbered amplicons for both the 6 and 14-amplicon strategies was to increase chances of full genome amplification in case one reaction or some of the amplicons failed and/or to reduce genome fragment amplification bias.

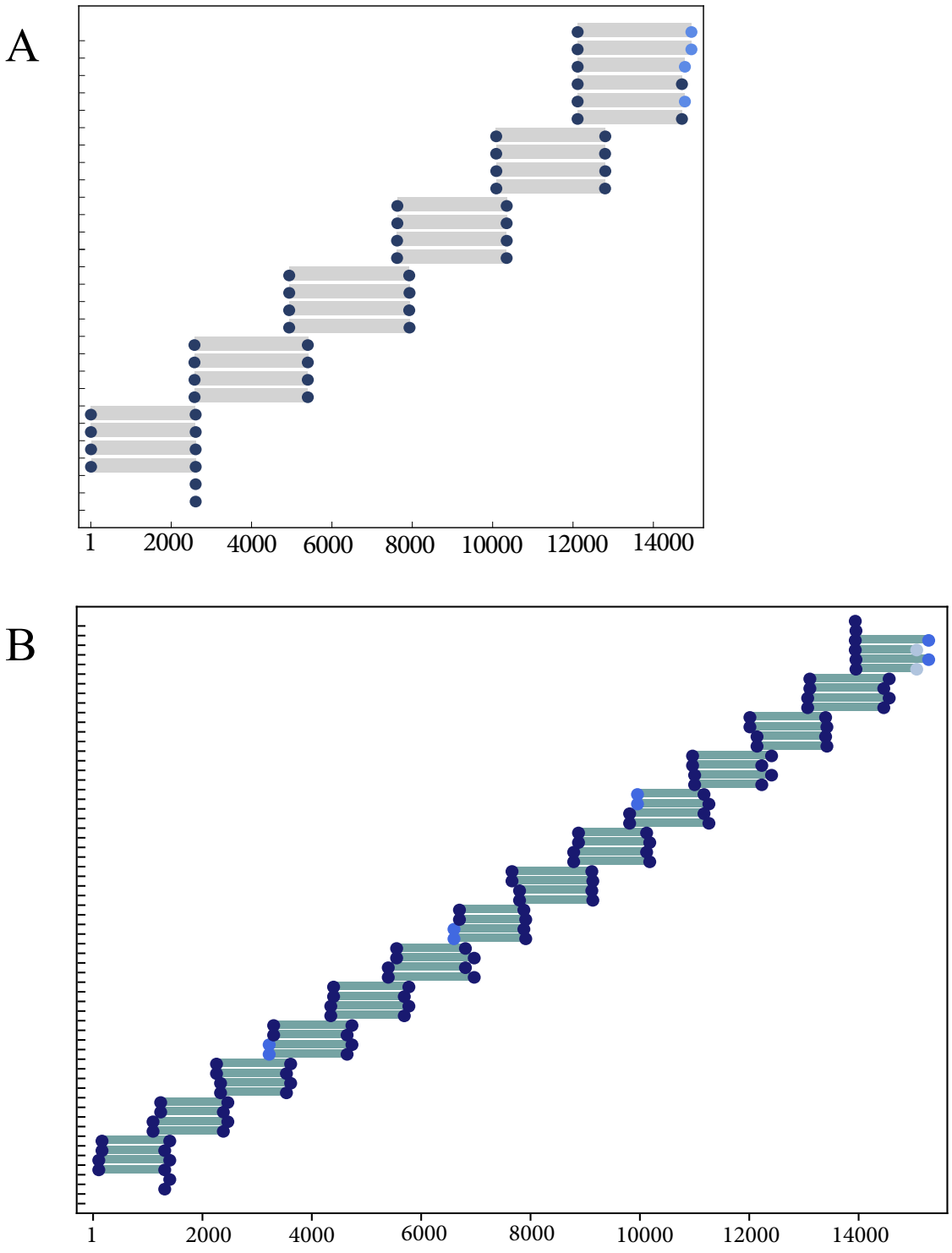


Figure 2.5: The RSV-A whole genome amplification strategies

The two panels show the (A) six and the (B) fourteen amplicons. For each panel the positions of primer targets for each amplicon are indicated by coloured circles (dark blue: perfect match, lighter blue: 1 mismatch, grey: 2 mismatches) while amplicon lengths are indicated by light green bars.

A reverse transcription primer mix was first prepared for each reaction. As the RSV genome molecule is negative stranded, the forward PCR primers were used for the reverse transcription. The respective forward primers for the different amplicons in a reaction were combined, e.g. forward primers for amplicons 1, 3, and 5 in reaction one and forward primers for amplicons 2, 4 and 6 in reaction two for the 6-amplicon two reaction strategy. A master mix was then prepared as shown in *Table 2.4*. Reagents 1-5 (enzyme mix) were premixed in a template free room and kept at -4ºC. Reagents 6-8, i.e. the template RNA, water and primers mix, were combined in the template addition hood, preheated at 65 ºC, and the containing tubes transferred to a frozen cooling block for one minute. The enzyme mix was then added to the primer-template mix and incubated at 50 ºC for 1 hour, 70 ºC for 15 minutes and finally held at 4 ºC until used for whole genome PCR amplification.

Table 2.4: Preparation of the WGS reverse transcription reaction mix

Description	1 reaction (µl)	Final concentration
5X First-Strand Buffer	4	1x
0.1 M DTT	1	0.005 M
RNase Inhibitor (40 units/µL)	1	2 units/µL
SuperScript™ III RT (200 units/µL)	1	10 units/µL
10 mM dNTP Mix	1	0.5 mM
Forward Primer Mix	1.9	0.38 pmol/µL
RNase free water	8.1	
Template RNA	2 µl	
Total reaction volume	**20 µl**	

Whole genome PCR primer mixes were similarly prepared as for the reverse transcription for each reaction. However, in this case both the forward and reverse primers were used. A reaction mix was prepared as shown in *Table 2.5*. The reaction mixtures were incubated at 98°C for 30 sec, and then 40 cycles of 98°C for 10 sec, 53°C for 30 sec, 72°C for 3.0 min, and final extension of 72°C for 10 min. Following PCR, an aliquot of the products for each reaction was run on a 0.6% agarose gel to check amplification success, and then pooled per sample in readiness for Illumina sequencing.

Table 2.5: Preparation of the whole genome amplification PCR reaction mix

	1 reaction (µl)	Final concentration
5 x Phusion HF buffer	5	1x
Mix of dNTPs *(10 mM each)*	0.5	0.2 mM
Phusion DNA Polymerase	0.25	
RNase-free water	12.35	
PCR primer mix,	1.9	
Template cDNA	5	
Total reaction volume	**25**	

2.8 Assembly of the short read NGS data

There are many genome assembly tools available and they differ greatly in terms of their performance (speed, scalability, hardware requirements and acceptance of newer read technologies) and in their final output (composition of assembled sequence) (Nagarajan and Pop 2013). This process of genome assembly is further complicated

by the different read lengths, read counts, single or paired reads, and error profiles that are produced by different (or even similar) NGS technologies. Assembling these short sequence reads without a reference genome is even more challenging as it is quite difficult to tell what's real, what's missing, and what's an experimental artefact (Baker 2012).

For this project, a set of four sequence assemblers in *Table 2.6* (not in any particular order) were evaluated with a few of our NGS reads. The assembly scaffolding tools SGA, SOPRA and SSPACE were also evaluated and their selection was based on being the best scaffolders from an analysis by Hunt *et al.* (Hunt *et al.* 2014). It was quite helpful that there were partial G gene Sanger dideoxy sequences for the samples that had undergone WGS for comparison.

Table 2.6: Genome assemblers tested on Kilifi RSV-A short read data

Assembler	Reason	Reference
SPAdes	Had been used by our group from a previous analysis (Agoti *et al.* 2015a).	(Bankevich *et al.* 2012)
Viral-ngs pipeline	A neat and easy to use pipeline for virus genomes assemblies, and previously used in the assembly of Ebola Virus genomes (Park *et al.* 2015).	https://viral-ngs.readthedocs.io/en/latest/
MaSuRCA	Best assembler from GAGE-B [Genome Assembly Gold-standard Evaluation for Bacteria] (Magoc *et al.* 2013)	(Zimin *et al.* 2013)
Vicuna	Targeted for diverse viral populations.	(Yang *et al.* 2012)

For the same dataset, varied output was observed in some of the assemblies evaluated e.g. spurious insertions with MaSuRCA (*Figure 2.6A*), with Vicuna (*Figure 2.6B*) and with SPAdes (*Figure 2.6C*) or missing 72nt duplication for known ON1 viruses with viral-ngs (*Figure 2.6D*). For viral-ngs, it is thought that the duplicate reads removal step might have over-filtered reads resulting in ON1 virus assemblies without the duplication. As viral-ngs and SPAdes had the least number of spurious insertions and deletions, and highest agreement with the previously sequenced Sanger dideoxy G gene assemblies, the two were retained for assembling all the short reads in this project and picked the best assembly for each sample from either based on the assembler with the longest genome fraction, best agreement with sanger G gene sequences, and no spurious insertions or deletions. No additional benefit was realized with the scaffolding tools on the draft genome assemblies hence they were not used further.

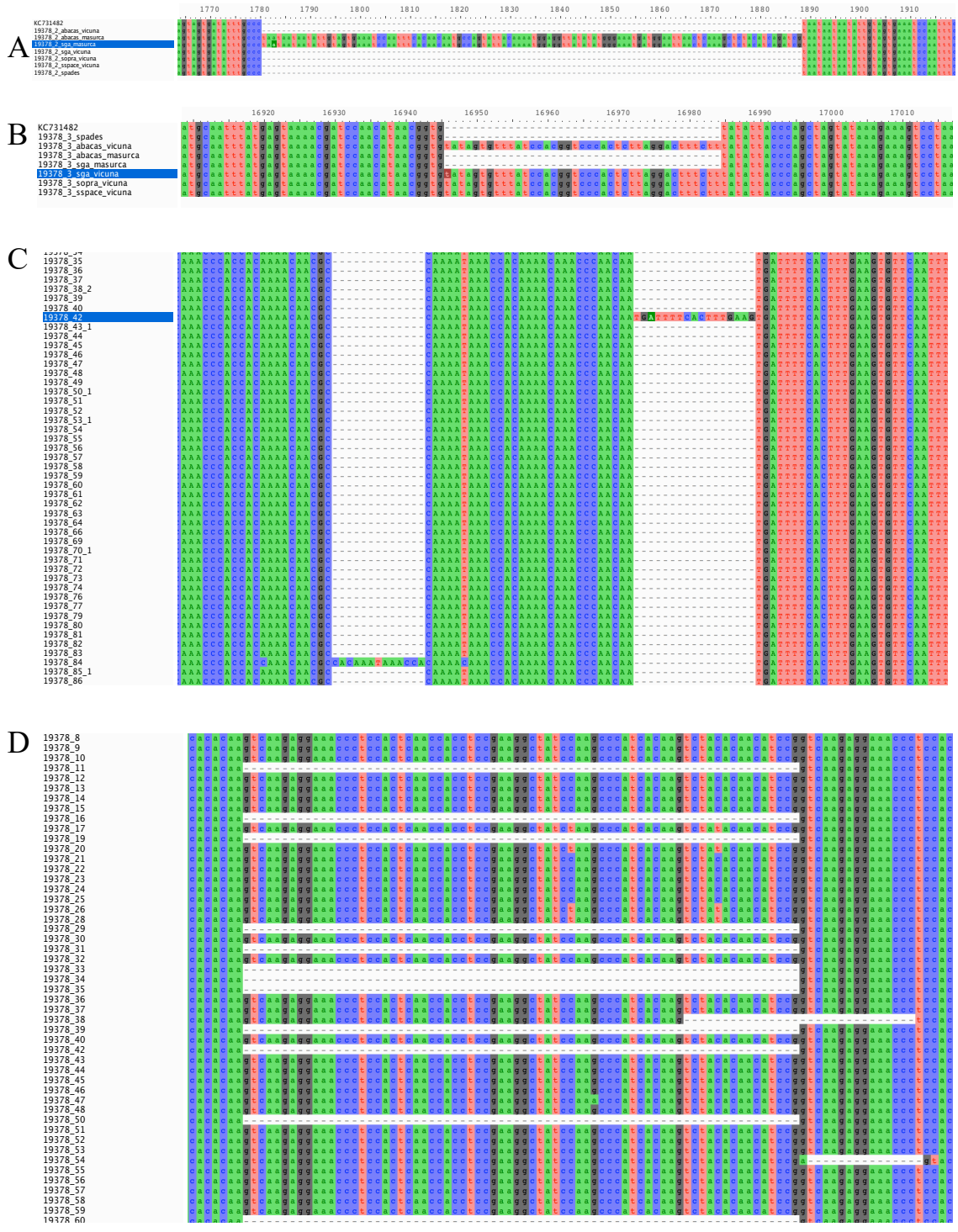


Figure 2.6: Evaluation of genome assemblers on Kilifi RSV-A datasets

2.9 Estimating the number of local variant introductions

To differentiate between local variants arising from a recent introduction and imported variants with greater genetic divergence than is expected from local diversification, we used a pragmatic criterion previously described by (Agoti *et al.* 2015b).

A variant is a virus (or a group of viruses) within a genotype that possesses $\geq N_d$ nucleotide differences compared to other viruses, where:

$$N_d = L_a S_r T_e$$

i.e. the product of the length of the genomic region analysed (L_a), estimated substitution rate for that region (S_r), and time elapsed (T_e).

We then calculated:

$$N_d / L_a \,.\, 100$$

to get the % nucleotide difference which was then used in usearch (Edgar 2010) to find the number of clusters/variants within a sequence dataset.

CHAPTER THREE

3 Molecular epidemiological, clinical and demographic characteristics of RSV-A genotype ON1 in Kilifi: analysis of G gene sequences

3.1 Background

Of the 11 proteins encoded by the RSV genome, the attachment glycoprotein (G) is the most variable and shown to accumulate considerable amino acid changes over time (Cane and Pringle 1995). RSV is classified into two groups, RSV-A and RSV-B (Mufson *et al.* 1985), with each group divided into genotypes (Peret *et al.* 1998) and these further characterized into variants (Agoti *et al.* 2015b). Globally, viruses belonging to different RSV groups, genotypes and variants often co-circulate in epidemics (Agoti *et al.* 2015b; Otieno *et al.* 2016). The phenomenon of reinfection and the difficulty in developing a vaccine may in part be due to the antigenic diversity and variability in the virus, and specifically the G gene (Cane 2001).

Two novel RSV genotypes with large duplications within the attachment G glycoprotein have been detected globally. In 1999, the BA genotype was detected in Buenos Aires, Argentina, based on a 60-nucleotide duplication within the C-terminal region of the G gene (Trento *et al.* 2003). The BA variant subsequently spread rapidly throughout the world becoming the predominant group B genotype and in some regions replacing all previous circulating RSV-B genotypes (Trento *et al.* 2010). More recently in December 2010, genotype ON1 with a 72-nucleotide duplication, also within the C-terminal region of the G gene, was detected in Ontario Canada (Eshaghi *et al.* 2012). Similarly, viruses belonging to this genotype have rapidly spread and diversified globally (Prifert *et al.* 2013; Tsukagoshi *et al.* 2013; Valley-

Omar *et al.* 2013; Agoti *et al.* 2014b; Auksornkitti *et al.* 2014; Pierangeli, Trotta and Scagnolari 2014; Avadhanula *et al.* 2015; Duvvuri *et al.* 2015). Such emergent genotypes appear to have a fitness advantage over preceding genotypes of the same RSV group (Hotard *et al.* 2015). Whether the potential fitness increase is associated with increased severity and immune evasion (with potential vaccine modality implications) is of considerable public health interest.

In February 2012, genotype ON1 was detected in Kilifi County, coastal Kenya, through routine partial G-gene sequencing. The temporal progression of RSV genotypes can be followed directly because of the distinctive tags (the duplications), providing a unique opportunity to understand more about the introduction, spread, severity and related selection processes (including immune evasion) for RSV, with the potential to lead into understanding the nature of emergence of novel virus variants. In this thesis chapter, an in-depth analysis of the epidemiological, clinical and sequence data of RSV-A viruses detected over five RSV seasons (2010/2011 to 2014/2015) was undertaken, and this includes the period from the initial detection of this novel genotype within Kilifi. This work used partial G gene sequences only with later work (Chapter Four) analyzing the WGS. Cases were diagnosed through paediatric pneumonia surveillance (RSV IP study) at the Kilifi County Hospital (KCH).

The work described in this chapter has been published and can be found from the following link; https://wwwnc.cdc.gov/eid/article/23/2/16-1149_article.

3.2 Aims of the Chapter

Using the immediately available and routinely sequenced partial G gene sequences together with patient clinical data, the aim of the chapter was to determine (1.) how new RSV variants get into and spread within a community with regard to the pace of entry [gradual or rapid] and nature of introductions [single or multiple variants], and (2.) the demographic and clinical impacts of such new variants.

3.3 Methods

Study location and population

The study was undertaken in Kilifi County, coastal Kenya, and is part of surveillance aimed at understanding the epidemiology and disease burden of RSV-associated pneumonia cases in this region (Nokes *et al.* 2009). Respiratory swab samples (combined nasopharyngeal and oropharyngeal) were collected between September 2010 and August 2015 from children aged 1 day to less than 5 years admitted to Kilifi County Hospital (KCH) presenting with syndromically defined severe or very severe pneumonia (referred to here as lower respiratory tract infections, LRTI) (Nokes *et al.* 2009).

Study samples and laboratory procedures

All specimens were screened for RSV by two methods (Nokes *et al.* 2004, 2009; Hammitt *et al.* 2011). Raw samples were tested for RSV antigen by Immunofluorescence Antibody Test (IFAT, Chemicon). Viral RNA was extracted from respiratory samples using QIAamp Viral RNA Mini Kit (QIAGEN) and tested for RSV (differentiating groups A and B) by multiplex real-time reverse transcriptase polymerase chain reaction (RT-PCR). All RSV positive samples tested positive by

either assay were taken forward for processing. Additionally, a small number of RSV negative samples were processed in parallel.

The viral RNA was reverse transcribed into cDNA using the Omniscript RT Kit (QIAGEN). The cDNA was then amplified with primers targeting the G ectodomain region (Scott *et al.* 2004; Agoti *et al.* 2012), and sequencing of the amplicons performed using BigDye v3.1 Chemistry on an ABI 3130xl. Sequence reads were assembled into contigs using Sequencher® v5.0.1 (Gene Codes Corporation, USA). The sequences analyzed here have been deposited in GenBank under the accession numbers KX453303 - KX453534. Previously reported sequences from Kilifi added to this analysis had been deposited in GenBank with accession numbers KF587911–KF588014 (Agoti *et al.* 2014b).

Global comparison dataset

To determine the relatedness of the Kilifi viruses to those circulating around the world and thereby understand their global context, all RSV-A G-gene sequences deposited in GenBank as of 19th January 2016 of length 241 to 687 nucleotides collected between 2010 and 2015 (inclusive) were downloaded. A total of 995 sequences from 24 countries were used in this analysis. For the whole dataset and for some of the countries, sequences were further binned by calendar year for temporal analysis. Unique sequences, identified as sequences that differ by at least one nucleotide from any other sequence over the sequenced region, were sub-sampled by epidemic season (Kilifi only) or per calendar year.

Sequence alignments and diversity analysis

All the sequences, from Kilifi and the global dataset, were collated and aligned using MAFFT alignment software v7.272 (Katoh and Standley 2013). Nucleotide and amino acid variability was calculated using MEGA v6.06 (Tamura *et al.* 2013).

Phylogenetic analyses

Maximum-Likelihood (ML) phylogenetic trees were inferred by MEGA 6.06 under the general time reversible (GTR) model with gamma (G) distributed among site rate heterogeneity model (Tamura *et al.* 2013). The GTR+G model was the best substitution model as determined by IQ-TREE v.1.4.2 (Chernomor, von Haeseler and Minh 2016). Bootstrapping with 1,000 iterations was implemented to evaluate branch support of the phylogenetic clusters. RSV-A genotypes were assigned as previously determined by *Peret et al* (Peret *et al.* 1998) and *Eshaghi et al* (Eshaghi *et al.* 2012). To position the genotype ON1 viruses in the global context, ON1 lineages were examined as recently assigned by *Duvvuri et al* (Duvvuri *et al.* 2015).

RSV-A variants analysis

The number of genotype GA2 and ON1 variants circulating in Kilifi and globally were calculated using a recently developed pragmatic criterion (Agoti *et al.* 2015b; Otieno *et al.* 2016), see the Methods chapter for specific details. Briefly, a variant is a virus or a group of viruses within a genotype that possesses ≥4 nucleotide differences (N_d) in the G ectodomain region compared to other viruses. This analysis was done using usearch v8.1.1861 (Edgar 2010).

Protein substitution and selection analysis

The prediction of *N*-glycosylation sites was performed using the NetNGlyc 1.0 server (Gupta, Jung and Brunak 2004). Only the default Asn-X-Ser/Thr sequon (where X is not proline) was considered for prediction. Patterns of change in amino acids were also analyzed using python scripts. Finally, potential positively selected and co-evolving sites were analyzed using the Datamonkey server (http://www.datamonkey.org/). For positive selection analysis, three methods were used; SLAC (Single Likelihood Ancestor Counting), FEL (Fixed Effects Likelihood) and MEME (Mixed Effects Model for Evolution).

Statistical analyses

Associations between demographic, clinical or outcome variables and RSV genotypes for all RSV positive severe and very severe pneumonia cases were explored using logistic regression and computing odds ratios using STATA v13 (StataCorp, Texas).

3.4 Results

3.4.1 ***RSV group and genotype temporal patterns***

Over the five RSV epidemics examined (2010/2011 to 2014/2015), 4,010 samples were collected from eligible children, 3,561 (88.8%) were tested for RSV and 881 (24.7%) RSV positives identified, *Table 3.1*. Of these, 600 (68.1%) were RSV group A. Sequencing of the G-gene was successful for 442 (73.7%) samples. A further 41 sequences were available from samples that were negative by both IFAT and PCR or from patients with mild pneumonia (data for these cases were, however, not included in the clinical severity analysis). Hence a total 483 sequences were taken forward for phylogenetic analysis. The sequences ranged 618-690 nt in length corresponding to nucleotides 295-912 of the reference strain A2 (M74568).

Two RSV-A genotypes were identified to be circulating in Kilifi over the five epidemics: ON1 (n = 283, 58.6%) and GA2 (n = 200, 41.4%). The temporal prevalence of the total RSV, RSV-A and genotypes ON1 and GA2 is shown in *Figure 3.1* and *Table 3.1*. Rapid replacement of the previously circulating dominant GA2 by ON1 in Kilifi was observed, from a prevalence of 0% in the 2010/2011 epidemic to 67.4% in 2011/2012 when ON1 was first detected in Kilifi, and to 96.1% in the recent 2014/2015 epidemic. In addition, RSV-A predominated in three consecutive RSV epidemics from 2012/2013 to 2014/2015. Kilifi long-term RSV-A genotype patterns are shown in *Appendix 7.4*.

Table 3.1: Frequency of LRTI inpatient cases, samples tested, total RSV and RSV-A cases, and number sequenced over five successive epidemics (2010/2011 to 2014/2015) in Kilifi, Kenya

Epidemic season[*]	LRTI eligible cases	Samples tested (%[†])	Number (%[‡]) of RSV positive samples	Number (%[§]) of RSV-A samples	Number of RSV-A sequences[#]	Number (%[¶]) of ON1 sequences	Number (%[¶]) of GA2 sequences
2014/2015	876	866 (98.9)	203 (23.4)	133 (65.5)	128	123 (96.1)	5 (3.9)
2013/2014	722	576 (79.8)	124 (21.5)	74 (59.7)	68	58 (85.3)	10 (14.7)
2012/2013	659	563 (85.4)	142 (25.2)	107 (75.4)	87	71 (81.6)	16 (18.4)
2011/2012	814	718 (88.2)	151 (21.0)	54 (35.8)	46	31 (67.4)	15 (32.6)
2010/2011	939	838 (89.2)	261 (31.1)	232 (88.9)	154	0 (0)	154 (100)
Total	**4,010**	**3,561 (88.8)**	**881 (24.7)**	**600 (68.1)**	**483**	**283 (58.6)**	**200 (41.4)**

[*]Epidemic designated 1st September of one year to 31st August of the following year

[†]As a proportion of the eligible LRTI inpatient cases

[‡]As a proportion of the samples tested

[§]As a proportion of the RSV positive samples

[#]Includes 41 sequences from IFAT-ve/PCR-ve or mild pneumonia cases

[¶]As a proportion of the RSV-A sequences

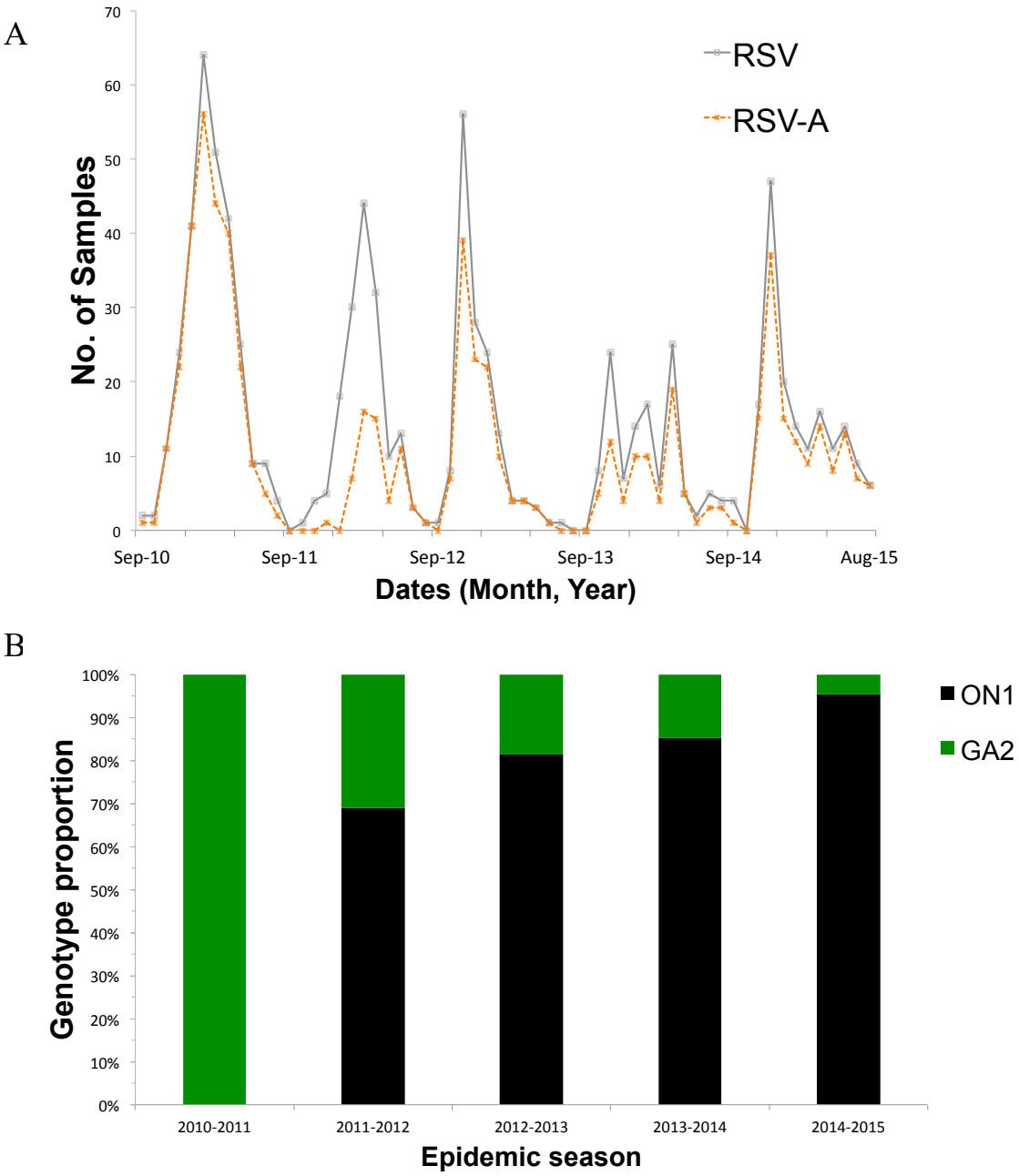


Figure 3.1: Circulating patterns of RSV in Kilifi, Kenya, September 2010 to August 2015. *1A*: Total RSV positive cases (grey continuous line) and typed RSV-A samples (dotted orange line) by month. *1B*: The proportion of RSV-A genotypes ON1 (black) and GA2 (green) per epidemic season. An RSV epidemic season is designated to start in September of one year until August of the following year. Unusually, for the last 3 seasons group A represents the vast majority of all RSV cases.

3.4.2 ***Demographic and clinical impact of ON1 in Kilifi***

To investigate the demographic and clinical impact of RSV-A genotype ON1 in Kilifi, a comparison was made of the proportions of GA2 and ON1 cases by gender, age, clinical features of cough, difficulty in breathing, chest wall indrawing, inability to drink, hypoxia, prostration/consciousness, pneumonia status (severe or very severe pneumonia), length of hospital stay, and death at the hospital (*Table 3.2*). The proportions of both genotypes were very similar for the demographic and clinical characteristics analyzed. Only the proportion of ON1 cases presenting with inability to feed was more than double that for GA2 cases (18.9% versus 8.8%), and this difference was found significant by logistic regression [OR 2.40 95% CI 1.31 – 4.36], *Table 3.3*, even after adjusting for age (categorical; <1 yr. and >=1 yr.) [OR 2.44 95% CI 1.31 – 4.57]. However, the proportion with very severe pneumonia was no higher in ON1 than in GA2 cases (OR 0.89, 95% CI 0.57 – 1.39).

Table 3.2: Demographic and clinical characteristics of RSV-A genotypes ON1 and GA2 in cases of severe or very severe pneumonia aged 1 day to less than 5 years admitted to Kilifi County Hospital September 2010 through August 2015.

		ON1	GA2	Total
		Number (%)	Number (%)	Number (%)
Age	< 1 year	223 (85.4)	158(87.3)	381 (86.2)
	> 1 year	38 (14.6)	23 (12.7)	61 (13.8)
Gender	Female	118 (45.2)	70 (38.6)	188(42.5)
	Male	143 (54.8)	111 (61.3)	254 (57.5)
Cough	No	5 (1.9)	4 (2.2)	9 (2.0)
	Yes	256 (98.1)	177 (97.8)	433 (98.0)
Breathing difficulty	No	15 (5.8)	3 (1.7)	18 (4.1)
	Yes	246 (94.3)	178 (98.3)	423 (95.9)
Chest wall indrawing	No	6 (2.3)	6 (3.3)	12 (2.7)
	Yes	255 (97.7)	175 (96.7)	430 (97.3)
Inability to feed	No	211 (81.2)	165 (91.2)	376 (85.3)
	Yes	49 (18.9)	16 (8.8)	65 (14.7)
Oxygen saturation	Above 90%	214 (82.0)	144 (79.6)	358 (81.0)
	Below 90%	47 (18.0)	37 (20.4)	84 (19.0)
Prostration/Consciousness	No	240 (92.0)	172 (95)	412 (93.2)
	Yes	21 (8.0)	9 (5.0)	30 (6.8)
Pneumonia status*	Severe*	203 (77.8)	137 (75.7)	340 (76.9)
	Very severe†	58 (22.2)	44 (24.3)	102 (23.1)
Hospital stay	1-4 days	160 (62.0)	101 (55.8)	261 (59.4)
	>4 days	98 (38.0)	80 (44.2)	178 (40.5)
Outcome	Alive	250 (96.9)	177 (97.8)	427 (97.3)
	Dead	8 (3.1)	4 (2.2)	12 (2.7)

*The denominator was 442 samples but for some of the features it ranged from 439 to 442

†(Cough OR difficulty in breathing) AND chest wall indrawing

‡(Cough OR difficulty in breathing) AND (hypoxic OR prostrate/unconscious)

Table 3.3: Clinical severity comparison between RSV-A genotype ON1 and GA2 cases of severe or very sever pneumonia aged 1 day to less than 5 years admitted to Kilifi County Hospital September 2010 through August 2015

		Unadjusted OR (95% Confidence Interval)	**p-value**
Age	< 1 year	0.85 (0.49 – 1.49)	0.579
Gender	Male	0.76 (0.52 – 1.12)	0.172
Clinical Presentation	Cough	1.16 (0.31 – 4.37)	0.830
	Breathing difficulty	0.28 (0.08 – 0.97)	0.045
	Chest wall indrawing	1.46 (0.46 – 4.59)	0.520
	Inability to feed	2.40 (1.31 – 4.36)	0.004
	Oxygen saturation <90%	0.86 (0.53 – 1.38)	0.521
	Prostration/Consciousness	1.67 (0.75 – 3.74)	0.211
Pneumonia status	Very severe pneumonia	0.89 (0.57 – 1.39)	0.609
Hospital stay	>4 days	0.77 (0.53 – 1.14)	0.192
Outcome	Dead	1.416 (0.42 – 4.78)	0.575

*The denominator was 442 samples but for some of the features it ranged from 439 to 442

3.4.3 ***Local Kilifi ON1 lineages as determined by the G-gene***

The ML tree in *Figure 3.2* shows the clustering of unique genotype ON1 sequences in Kilifi. Two genotype ON1 lineages initially defined by Duvvuri *et al* were shown to be circulating in Kilifi, i.e. lineages ON1 [1.1] and ON1 [1.3] (Duvvuri *et al.* 2015). Of these two lineages, ON1 [1.3] was the most prevalent in 2011/2012 and 2012/2013. However, a potential new lineage, denoted here as ON1 [1.4], clustered away from ON1 [1.3], and seemed to have recently arisen comprising of sequences from the 2013/2014 and 2014/2015 epidemics. The genetic divergence (p-distance) between ON1 [1.4] and the other ON1 lineages identified in Kilifi ranged between 0.013 – 0.045, which was similar to the genetic distances between the previously defined ON1 lineages (Duvvuri *et al.* 2015).

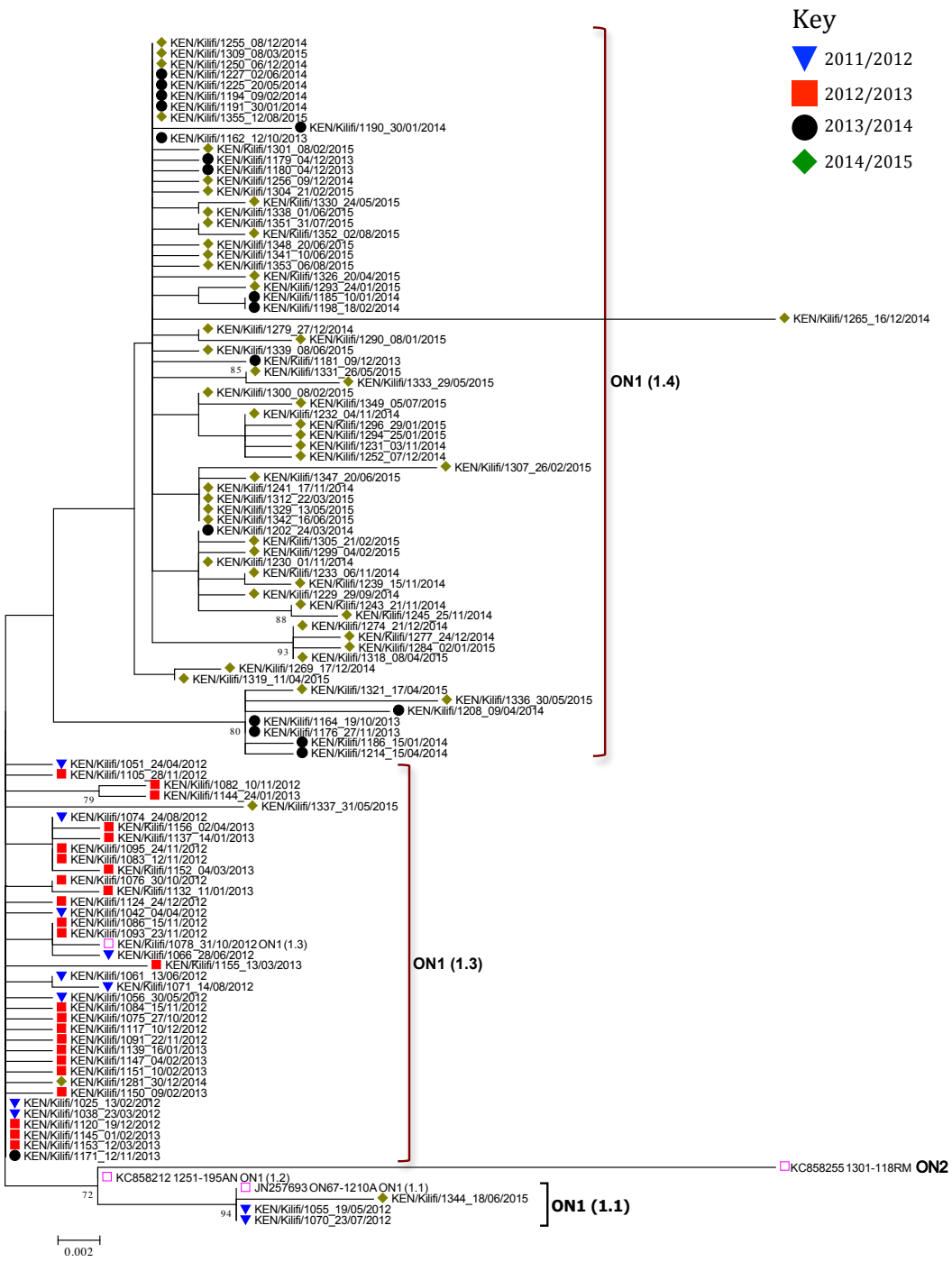


Figure 3.2: An unrooted ML phylogenetic tree of unique genotype ON1 G-gene ectodomain sequences from Kilifi, Kenya, 2012 to 2015.

The taxa are colour coded by the epidemic season of detection as shown by the key, and the names represent "KEN/Kilifi/serialno_date of collection". Note that although the study detected RSV ON1 in the epidemic season 2011-12, the first ON1 cases were in 2012. Lineages ON1 [1.1] – ON1 [1.3] initially assigned by Duvvuri *et al*. are shown in bold in addition to the newly assigned lineage ON1 [1.4] identified in Kilifi.

3.4.4 ***RSV-A introduction and persistence patterns***

A total of 66 RSV-A variants were detected over the whole surveillance period in Kilifi, *Figure 3.3*, where a variant is defined from Usearch using $N_d \geq 4$ nucleotides difference. The variants comprised of between 1-82 sequences, with 39 of the 66 variants (59.1%) being singletons. Most variants did not persist between epidemics (69.7%; 46/66). However, fourteen variants persisted for two consecutive seasons, one for four consecutive seasons and five for two non-consecutive seasons. Therefore, the number of variants when accumulated by epidemic increases to 86. The number of GA2 variants was observed to decline consistently over time, i.e. from seventeen variants in 2010/2011 (before ON1 arrived) to only four variants in 2014/2015. In contrast, the number of ON1 variants assigned remained at five variants between 2011/2012 and 2012/2013 before increasing to eight variants in 2013/2014, then rising markedly to 25 variants in 2014/2015.

Variant	Genotype	Epidemic (Number of variants per epidemic)				
		2010/2011	2011/2012	2012/2013	2013/2014	2014/2015
		17	15	13	12	29
1	ON1					
2	ON1					
3	ON1					
4	ON1					
5	ON1					
6	ON1					
7	ON1					
8	ON1					
9	ON1					
10	ON1					
11	ON1					
12	ON1					
13	ON1					
14	ON1					
15	ON1					
16	ON1					
17	ON1					
18	ON1					
19	ON1					
20	ON1					
21	ON1					
22	ON1					
23	ON1					
24	ON1					
25	ON1					
26	ON1					
27	ON1					
28	ON1					
29	ON1					
30	ON1					
31	ON1					
32	ON1					
33	ON1					
34	ON1					
35	GA2					
36	GA2					
37	GA2					
38	GA2					
39	GA2					
40	GA2					
41	GA2					
42	GA2					
43	GA2					
44	GA2					
45	GA2					
46	GA2					
47	GA2					
48	GA2					
49	GA2					
50	GA2					
51	GA2					
52	GA2					
53	GA2					
54	GA2					
55	GA2					
56	GA2					
57	GA2					
58	GA2					
59	GA2					
60	GA2					
61	GA2					
62	GA2					
63	GA2					
64	GA2					
65	GA2					
66	GA2					

Figure 3.3: Temporal occurrence of RSV-A variants (rows) detected in Kilifi Kenya, 2010/2011 to 2014/2015.

A variant was defined as a group of viruses (where group includes a singleton) within a genotype that possesses ≥4 nucleotide differences in the G ectodomain region compared to other viruses (see Methods). Genotypes are shown in different colours: GA2 (dark blue) and ON1 (green).

3.4.5 ***N-glycosylation patterns within the G-gene***

Seven codon sites were predicted to be *N*-glycosylated within the G protein for the Kilifi sequences: four sites for genotype ON1 viruses (codons 103, 135, 237 and 318) and six sites for genotype GA2 viruses (codons 103, 135, 237, 251, 273 and 294). However, none of the potential *N*-glycosylation sites occurred within the 72-nucleotide duplication region (144 nucleotides; codons 260-307) of the ON1 viruses. Interestingly for GA2 viruses, sites 237 and 273 were mutually exclusive such that a virus belonging to this genotype had either one of these sites potentially *N*-glycosylated but not both, *Figure 3.4*.

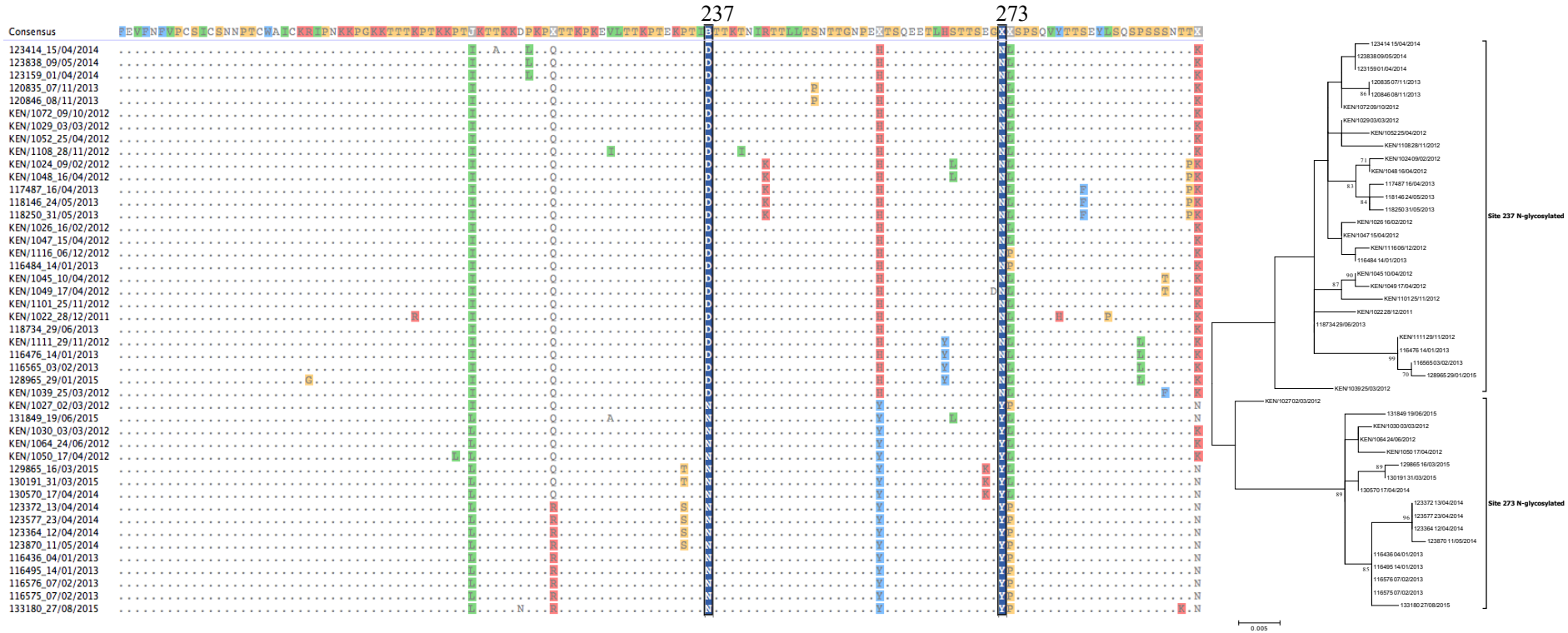


Figure 3.4: Differentiation of GA2 viruses in Kilifi based on N-glycosylation patterns.

Amino acid alignment and ML tree highlighting the two codon sites (237 and 273) on the G protein that define two groups of GA2 viruses in Kilifi based on *N*-glycosylation patterns Samples are from Kilifi County Hospital admissions aged 1 day to under 5 years, 2012-2015.

3.4.6 ***G-gene nucleotide and amino acid variability***

The nucleotide and amino acid variability over the four seasons is shown in *Table 3.4*. Amino acid substitutions over the sequenced portion of the G protein are shown in *Figure 3.5A*. Two codon positions possessed amino acid substitutions that distinguished between ON1 (232G, 253K) and GA2 (232E, 253T) viruses. In addition, the new ON1 [1.4] lineage viruses seem to have fixed a threonine (I/T136T) and acquired a unique substitution (P206Q) that distinguishes them from the other ON1 lineages.

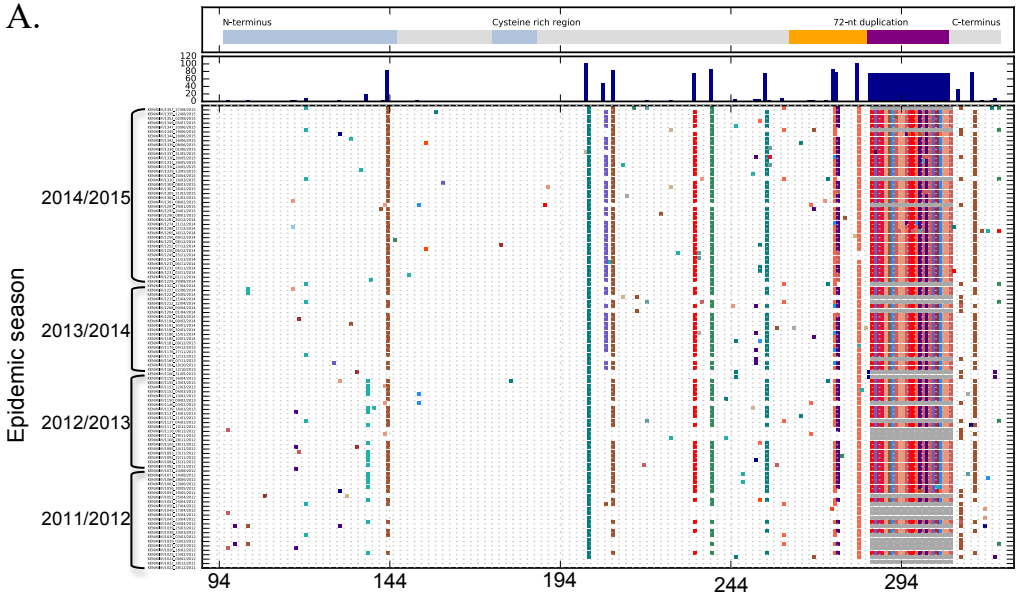
Table 3.4: G protein nucleotide and amino acid variability in RSV-A genotypes identified in Kilifi, Kenya, 2010/2011 to 2014/2015

Epidemic season*	All RSV-A		ON1		GA2	
	Nucleotide	Amino acid	Nucleotide	Amino acid	Nucleotide	Amino acid
2014/2015	0.88	1.40	0.68	1.10	2.02	4.12
2013/2014	1.24	2.07	0.55	0.75	1.72	3.29
2012/2013	1.21	2.14	0.46	0.58	1.69	3.13
2011/2012	1.45	2.99	0.43	1.02	1.56	2.91
2010/2011	0.66	0.99	-	-	0.66	0.99
Total*	**2.04**	**3.41**	**0.85**	**1.23**	**1.01**	**1.71**

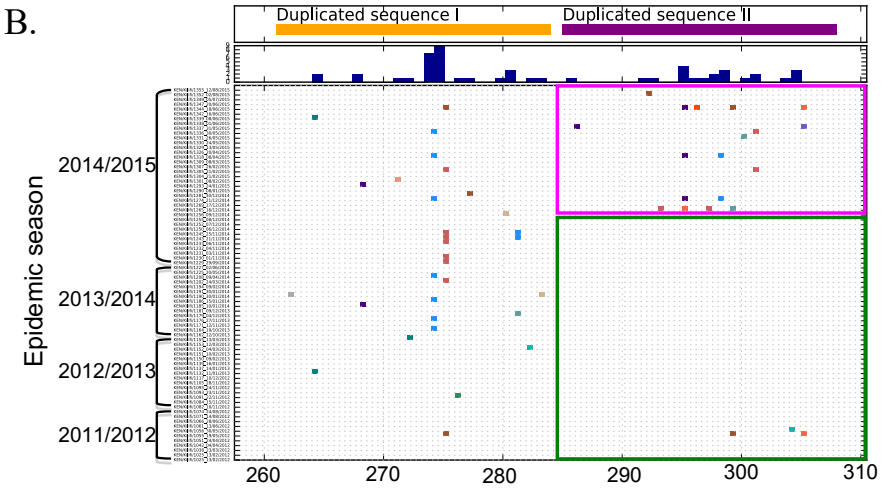
*Epidemic designated 1st September of one year to 31st August of the following year

Figure 3.5B and *3.5C* illustrate amino acid substitutions within the duplication region of the Kilifi ON1 viruses. The first set of 72 nucleotides was designated as "duplication sequence I" and the second set as "duplication sequence II". Within this region, it was observed that over the three seasonal epidemics of 2011/2012 to 2013/2014 epidemics and in early (Sept-Nov) 2014/2015 epidemic, amino acid substitutions only occurred within the duplicated sequence I except for three substitutions in two viruses that occurred within the duplicated sequence II. From December 2014 of the 2014/2015 epidemic, however, numerous substitutions were detected in duplicated sequence II with two adjacent and corresponding positions between the duplicated sequences I and II acquiring similar amino acid substitutions, i.e. Y273H and Y297H; P274L/S and P298L/R. Furthermore, sites 273 and 297 were detected to be co-evolving from the Spidermonkey analysis. However, only one ON1 codon site (251) was identified to be positively selected with $p<0.05$ by at least one method (SLAC, FEL and MEME).

A.



B.



C.

Duplicated sequence I	Major AA (alternative, freq.)	Duplicated sequence II	Major AA (alternative, freq.)
260	S	284	G
261	Q (-, 1)	285	Q (P, 1)
262	E	286	E
263	E (K, 2)	287	E
264	T	288	T
265	L	289	L
266	H	290	H
267	S (P, 2)	291	S (L, 1)
268	T	292	T (S, 1)
269	T	293	T
270	S (T, 1)	294	S (P, 3, Y, 1)
271	E (K, 1)	295	E (D, 1)
272	G	296	G (S, 1)
273	**Y (H, 7)**	**297**	**Y (H, 2)**
274	**P (S, 7, L, 2)**	**298**	**P (L, 2, R, 1)**
275	S (N, 1)	299	S (R, 1)
276	P (L, 1)	300	P (S, 2)
277	S	301	S
278	Q	302	Q
279	V (A, 1)	303	V (I, 1)
280	Y (H, 2, C, 1)	304	H (Y, 2, Q, 1)
281	T (I, 1)	305	T
282	T (A, 1)	306	T
283	S	307	S

Figure 3.5: Amino acid substitutions in RSV-A G-protein for sequences isolated in Kilifi Kenya from season 2011/2012 to 2014/2015.

All unique protein sequences per epidemic were collated, aligned and the amino acid differences from the earliest sequence determined and marked with vertical coloured bars, with the substituted amino acid residue colour coded as shown by the key between panels *A* and *B*. Panel *A* shows the full aligned AA sequence inferred from the G gene sequences (ON1 and GA2) whereas Panel *B* (ON1 only) focuses on the region of the ON1 duplication. The positions shown at the bottom of panels *A* and *B* are relative to the first amino acid of the regions analyzed, i.e. from amino acid positions 94 and 260, respectively, of the reference strain A2 [Ref: M74568]. Indicated at the top of these panels are the functional domains of the G protein (*A*) and the 72-nucleotide duplication of genotype ON1 (*B*; duplicated sequence I in "orange" and duplicated sequence II in "purple"). Below this, the histogram indicates the total number of changes at each position. The green and pink rectangles in Panel *B* represent periods of minimal and numerous substitutions, respectively, within the duplicated sequence II. Panel *C* shows concurrent AA positions within the duplicated sequences I and II, and the respective amino acid substitutions (numbering similar to positions in Panel *B*). In bold are concurrent positions with similar amino acid substitutions.

3.4.7 ***The global dynamics of ON1 vs BA variants***

Using global datasets for genotypes ON1 and BA, we compared the temporal detection of RSV variants within each of these genotypes over the first five and ten years, respectively, from initial detection; *Table 3.5*. There was an explosion of new ON1 variants globally from eight variants in 2011, to 78 variants in 2012, and to 153 variants in 2013. However, there was a decrease in the number of ON1 variants in 2014 and 2015, which corresponded with a substantial decrease in both the number of ON1 sequences available in GenBank and the countries that have deposited sequences from these years. On the other hand, the number of BA variants followed a stepwise or punctuated pattern whereby the number of variants was stable between one to six variants from 1996-2001, then drastically increased to and stabilized between 20 to 30 variants from 2002-2004, and again sharply rising to 82 variants in 2005. At the country level, *Table 3.6*, the rapid rise in the number of ON1 variants detected in Kilifi was similarly observed in Philippines and Germany while the number of BA variants detected in some of the countries sampled remained relatively stable over time.

Table 3.5: Frequency of global genotypes BA and ON1 variants detected, by calendar year, 1998-2015

BA*				ON1			
Year	Variants[$]	Sequences	Countries	Year	Variants	Sequences	Countries
1998	2	4	2	2010	1	5	2
1999	3	11	4	2011	8	57	7
2000	2	2	1	2012	78	368	19
2001	6	23	4	2013	153	432	15
2002	20	51	9	2014	59	256	8
2003	30	60	10	2015	10	81	2
2004	30	144	12				
2005	82	290	13				
2006	101	205	16				

*Data extracted from GenBank on 19th January 2016, length 241 to 687 bases.
[$]Variant defined as $N_d \geq 4$

Table 3.6: Select country specific frequency of genotypes BA and ON1 variants, by calendar year, 1999-2015

Genotype	**Country**	**Year and variants***							
		2010	**2011**	**2012**	**2013**	**2014**	**2015**		
ON1	Philippines	-	-	7	29	-	-		
	Germany	-	-	5	26	-	-		
	Japan	-	-	5	9	7	1		
	Spain	-	-	5	7	14	-		
BA		**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
	Belgium	2	-	5	5	5	5	8	9
	Argentina	1	-	-	6	8	5	-	
	Japan	-	-	1	2	4	3	9	6

*Data extracted from GenBank on 19th January 2016, length 241 to 687 bases.

3.5 Discussion and Conclusions

In this chapter, we provide a detailed analysis on the spread and the associated demographic, clinical and evolutionary characteristics of the novel RSV-A genotype ON1 in Kilifi. ON1 was first detected in Kilifi in February 2012, and within that RSV epidemic (2011/2012) it displaced the previous dominant genotype GA2 by attaining a prevalence of 67%. Its dominance continued to rise to 96% within a span of only four epidemics. This rapid rate of replacement is unlike previous observations from the same location on the displacement of GA5 by GA2; it took GA2 about seven years to attain a similar prevalence of 95% (Otieno *et al.* 2016). ON1 seems to possess a greater fitness advantage over GA2 than GA2 over GA5. If such fitness is the result of immune evasion, there are potential implications for vaccines to deliver population level immunity via herd protection (Kinyanjui *et al.* 2015).

The present study found evidence for increased severity of cases of RSV ON1 relative to GA2, showing a higher risk for inability to feed. However, overall cases of very severe pneumonia were equally prevalent in both genotypes. The data, therefore, does not provide a strong indication of more severe disease arising from the ON1 variant. Duvvuri *et al.* (Duvvuri *et al.* 2015) reported significant association of ON1 with females, which was not evident in the present study. Yoshihara *et al* (Yoshihara *et al.* 2016) reported that ON1 ARI cases in Vietnam were significantly associated with clinically severe presentations of wheezing, tachypnoea and difficulty in breathing as compared to NA1 cases while Panayiotou *et al* (Panayiotou *et al.* 2014), on the contrary, reported that children infected with ON1 in Cyprus experienced significantly milder illness compared to infections with GA2. Some studies have reported no differences at all between genotypes in risk of clinical severity. The

discordant results may arise from methodological differences in analyses, clinical disease definitions and study designs, chance effects resulting from inadequate sample sizes, differences between viruses in different locations, or even host/environmental differences. Prospective studies specifically designed to evaluate virulence or clinical differences between genotypes may offer more reliable insight.

We noted some change in the alternation of RSV subgroup dominance pattern in Kilifi since the introduction of ON1 into this community. While replacing GA2, the presence of ON1 appeared to also exclude group B strains. RSV-A predominated over RSV-B in three consecutive epidemics from 2012/2013 to 2014/2015. The epidemics thereafter of 2015/2016, 2016/2017 and 2017/2017 were predominated by RSV-B, RSV-B and RSV-A, respectively. Previously, using data collected between 2002 and 2012 in Kilifi, RSV-A predominated in only up to two consecutive epidemics (Otieno *et al.* 2016). It is unclear whether the RSV-A predominance in the three consecutive epidemics was related to ON1, a general change in RSV epidemiological patterns or a chance occurrence.

Globally, the reported prevalence of ON1 varies from one location to another. In Ontario, Canada, where ON1 was first detected in December 2010, the prevalence of ON1 has remained stable at 11-13% [2011-2012] (Duvvuri *et al.* 2015). Other countries that have similarly reported ON1 prevalence rates <20% include South Africa [2012] (Pretorius *et al.* 2013) and China [2011-2013] (Yu *et al.* 2015). Reports from Italy [2011-2013] (Pierangeli, Trotta and Scagnolari 2014), South Korea [2011-2013] (Kim *et al.* 2014), USA [2011-2013] (Avadhanula *et al.* 2015), Malaysia [2011] (Khor *et al.* 2013), Japan [2012] (Tsukagoshi *et al.* 2013), Thailand [2010-

2011] (Auksornkitti *et al.* 2014), Latvia [2009-2012] (Balmaks *et al.* 2014) and Cyprus [2010-2013] (Panayiotou *et al.* 2014) indicate a range of ON1 prevalence of between 20% to 70%. An article from Argentina reported the detection of ON1 as the only (100%) RSV-A genotype in Buenos Aires in 2014 (Viegas, Goya and Mistchenko 2016). Even though the prevalence rates reported in these publications may not have been long after ON1 introductions into those countries, the varied prevalence suggests that while ON1 is rapidly spreading globally there could be host or ecological differences that determine RSV spread.

As shown here and in previous work from Kilifi (Agoti *et al.* 2015b; Otieno *et al.* 2016), RSV epidemics are comprised of multiple variants to suggest separate introductions into the community (as opposed to arising from diversification during the epidemic). In addition, these variants often do not persist between epidemics. This suggests that each year (a) variants generate local herd immunity leading to their demise hence requiring reintroductions or (b) that there is competition for seeding new seasonal epidemics from many invading variants and the preceding year variants lose out (perhaps on a chance basis or as stated above because they have less fitness due to variant specific immunity). However, the notion of multiple introductions needs to be demonstrated by genetic diversity and phylogenetic analyses using full genome data that should have a better resolution at distinguishing between virus variants.

RSV accumulates amino acid changes over time (Cane and Pringle 1995), and the same was similarly observed in the Kilifi ON1 viruses. It is of interest that in the first three RSV epidemic seasons of ON1 detection in Kilifi, amino acid substitutions were

restricted to the duplicated sequence I of the 72nt duplication save for three substitutions in two viruses that occurred within the duplicated sequence II. However, in the 2014/2015 epidemic an 'explosion' in amino acid substitutions was observed within the duplicated sequence II that coincided with a surge in the number of detected ON1 variants. In addition, there were similar amino acid substitutions in two adjacent and corresponding sites in the duplicated sequences I and II, with one set of these sites co-evolving. Considering that the 72nt duplication in ON1 viruses represents a longer attachment protein, this may offer more opportunities for variable changes, greater diversity and increased fitness over previous group A genotypes.

Two codon sites, 232 and 253, within the G protein region analyzed were found to distinguish between genotype ON1 and GA2 viruses. The amino acid change Glu-232-Val has been reported in RSV-A escape mutants that result in loss of reactivity to a specific monoclonal antibody (Cane and Pringle 1995). Further, a functional analysis of the 60-nucleotide duplication in BA strains has shown that the duplicated region in the G protein of these viruses augment their fitness (Hotard *et al.* 2015). A similar analysis has not been reported for ON1, but it is plausible that the 72-nucleotide duplication may contribute considerably in the increased fitness observed in. However, G-protein *N*-glycosylation seems to play no role in the increased fitness of ON1 as similar potential *N*-glycosylation codon sites were detected in both ON1 and GA2 and no additional *N*-glycosylation sites were detected within the ON1 duplication region.

The rapid diversification of ON1 observed in Kilifi seems to reflect rapid expansion at the global level. While sampling variability may play a role, and similar to varying

prevalence, there was variability in the diversification of ON1 viruses in different countries. The number of ON1 variants seemed stable in some countries (e.g. Japan) while expanding in others (e.g. Germany and Philippines). The temporal distribution of BA variants in different countries, however, was mostly stable. Comparisons in the temporal patterns of genotype BA and ON1 variants may highlight differences between the RSV group B and A viruses. Increased sampling and surveillance will help illuminate on whether such inter-genotypic and inter-group differences are due to ecological differences or variable sampling.

In conclusion, it is evident that genotype ON1 is not only rapidly spreading globally but also increasing in frequency. The result is the near exclusion of the previous dominant group A GA2 genotype in some locations. The implications of this apparent increased fitness of RSV-ON1 are yet to be resolved. There is some evidence for increased severity of the virus, but this is by no means clear or consistent across studies. Continued surveillance for cases together with collection of detailed standardized clinical data is warranted. The possibility exists that ON1 and other similar new RSV variants (e.g. the BA genotype) become dominant by evading host immunity. It is reasonable to assume this could lead to evasion of future vaccine induced protection, lessening the herd immunity potential of vaccination, similar to influenza A vaccines. Analysis of genome-wide substitutions between viruses belonging to incoming and outgoing genotypes has the potential of highlighting substitutions with likely functional and fitness implications.

CHAPTER FOUR

4 Whole genome evolutionary dynamics of RSV genotype ON1

4.1 Background

The single stranded negative-sense RSV genome is approximately 15.2Kb long and has 10 genes that encode 11 proteins (Collins and Melero 2011). However, the bulk of RSV molecular epidemiology studies to date have been (and still are) based on partial sequencing of the gene encoding the attachment G glycoprotein. This is because the G protein is the most variable RSV protein between and within viruses of the two co-circulating RSV groups, A and B (Johnson *et al.* 1987), in addition to being a key target of protective immune responses (Melero *et al.* 1997). In this chapter, the utility of WGS in gaining a better understanding of the molecular epidemiology of RSV was explored.

The G gene on its own has been shown to be insufficient in distinguishing between inpatient outbreak RSV strains isolated in a haematology-oncology and stem-cell transplant unit and outpatient epidemiologically-unrelated strains collected within the same time period (Zhu *et al.* 2017), and "who acquires infection from whom" in a household RSV transmission study (Munywoki *et al.* 2014). Further, an analysis by Agoti *et al. 2015* showed that viruses that were 100% identical in this genome region had substitutions in other parts of the genome (Agoti *et al.* 2015a). In addition to providing data on the consensus genome sequences, NGS provides information on low frequency minority sequence variants, which could be useful in understanding the evolutionary and transmission dynamics of the target organism (Bull *et al.* 2012; Henn *et al.* 2012; Grad *et al.* 2014; Ha Do *et al.* 2015; Rodriguez-Roche *et al.* 2016; Cobbin *et al.* 2017; Githinji *et al.* 2018). Therefore, WGS has potential for a more

detailed understanding of RSV molecular epidemiology, evolution, phylogeography, diagnostics and vaccine development.

The rapid replacement of the previously circulating RSV-A genotype GA2 by genotype ON1 since first detection in Kilifi in 2012 suggests some fitness advantage of ON1 over the GA2 viruses in Kilifi (Hotard *et al.* 2015; Otieno *et al.* 2017), and might be accompanied by other important genomic substitutions in ON1 viruses. Close observation of such a new virus genotype introduction over time provides an opportunity to better understand the transmission and evolutionary dynamics of the pathogen. This chapter presents WGS sequencing and analysis of 184 RSV-A genomes from Kilifi, the very first ON1 genomes from Africa, and advances the previous work on the patterns of introduction and persistence of the ON1 variant within this community that utilized partial G gene sequences (Agoti *et al.* 2014b; Otieno *et al.* 2017). The results of this analysis provide a higher resolution of the RSV genetic structure, spread and identification of variation that may be associated with molecular adaptation and apparent fitness advantages.

The work described in this chapter has been published and can be found from the following link; https://academic.oup.com/ve/article/4/2/vey027/5106641. Analysis files and scripts can be found on GitHub: https://github.com/jrotieno/Kilifi_ON1_genomes

4.2 Aims of the Chapter

The aims of this chapter were to utilize full length RSV-A genome sequences to identify genome-wide substitutions that differentiate between old and new RSV variants entering a community, characterize the genome regions where the

substitutions occur, and infer if such substitutions could impact on virus fitness, and to get a better understanding of the nature of emergence of the RSV-A ON1 genotype using WGS.

4.3 Methods

Study population

Two sets of samples were used in the current analysis; (i) samples collected as part of the RSV IP study from children (under 5 years of age) admitted to KCH presenting with syndromically defined severe or very severe pneumonia between September 2011 to August 2016 (Nokes *et al.* 2009; Otieno *et al.* 2017), and (ii) samples collected as part of the SPReD-KHDSS study from patients of all ages presenting at health facilities within the KHDSS with ARI between January to December 2016 (Scott *et al.* 2012; Nyiro *et al.* 2018).

RNA extraction and PCR amplification

All KCH admissions specimens had previously been screened for RSV (immunofluorescent antibody test, IFAT), RSV group (multiplex real-time polymerase chain reaction) and RSV-A genotype status (G gene amplification followed by Sanger sequencing), and partial G-gene sequencing results reported (Otieno *et al.* 2017), while the KHDSS samples were screened afresh using the same multiplex real-time PCR methods referred to as above. To pick samples proceeding to WGS, we selected (i) all the RSV-A positives from the KHDSS, (ii) all the GA2 positives from KCH, and (iii) a random subsample (50%) of the ON1 positives per epidemic from the KCH. Additionally, we targeted samples with real-time PCR cycle threshold (Ct) value < 30 based on the success rate from previous experience (Agoti *et al.* 2015a), with the exception of four test samples that were PCR negative or had

Ct>30. Viral RNA was extracted using QIAamp Viral RNA Mini Kit (QIAGEN, London, UK). Reverse transcription (RT) of RNA molecules and PCR amplification were performed with a six-amplicon, six-reaction strategy (Agoti *et al.* 2015a), or using a 6 or 14-amplicon strategy split into two reactions of three and seven amplicons, respectively for each (Chapter Two, *Figure 2.4*). Amplification success was confirmed by observing the expected PCR product size (1200-1500 bp) on 0.6% agarose gels. Amplicons from six or two reactions were pooled and purified for Illumina library preparation.

Illumina library construction and sequencing

The purified PCR products were quantified using Qubit fluorimeter 2.0 (Life Technologies) and normalized to 0.2 ng/μL. The normalized DNA was tagmented (a process of fragmentation and tagging) using the Nextera XT (Illumina, San Diego, USA) library prep kit as per the manufacturer's instructions. Indices were ligated to the tagmented DNA using the Nextera XT index kit (Illumina). The barcoded libraries were then purified using 0.65X Ampure Xp beads. Library quality control was carried out using the Agilent high sensitivity DNA kit on the Agilent 2100 Bioanalyzer (Agilent, Waldbronn, Germany) to confirm the expected size distributions and library quality. Each library was quantified using the Qubit fluorimeter 2.0 (Life Technologies), after which the libraries were normalized and pooled at equimolar concentrations. The pooled libraries were sequenced on either (i) Illumina HiSeq system using 2 x 250 bp paired-end (PE) sequencing at the Wellcome Trust Sanger Institute (UK), or (ii) Illumina MiSeq using 2 x 250 bp PE sequencing at the KEMRI-Wellcome Trust Research Programme (Kilifi, Kenya).

A preliminary quality check of the sequence reads was done using fastqc (Andrews 2010) with the output per batch aggregated and visualized by multiqc (Ewels *et al.* 2016). To determine the proportion of RSV and non-RSV reads in the samples used here, Kraken v0.10.6 (Wood and Salzberg 2014) was used with a pre-built Kraken database provided by the viral-ngs pipeline (Park *et al.* 2015, 2016) (downloaded in December, 2015; https://storage.googleapis.com/sabeti-public/meta_dbs/kraken_ercc_db_20160718 .tar.gz).

Depletion of human reads

Prior to deposition of the raw short reads into NCBI short read archive (SRA), datasets were depleted of human reads. The raw reads were mapped onto the human reference genome hg19 using bowtie2 (Langmead *et al.* 2009) while samtools (Li *et al.* 2009) was used to filter, sort and recover the unmapped (non-human) reads. The final reads are available in the NCBI BioProject database under the study accession PRJNA438443.

Genome assembly and coverage

Consensus genome assemblies were generated either using viral-ngs versions 1.18.0/1.19.0 (Park *et al.* 2015, 2016) and/or SPAdes version 3.10.1 (Bankevich *et al.* 2012), selecting the most complete assembly from either assemblers (more details in the Methods Chapter [Chapter 2.8]). The available Sanger G-gene sequences (Agoti *et al.* 2014b; Otieno *et al.* 2017) for these samples were additionally used to confirm agreement with the WGS assemblies. The genomes generated in this study are available in GenBank under accession numbers MH181878 - MH182061. The

genomes were aligned using MAFFT alignment software v7.305 (Katoh and Standley 2013) using the parameters '—localpair –maxiterate *1000*'.

To calculate and visualize depth of coverage, sample raw reads were mapped onto individual assemblies with BWA (Li and Durbin 2009), samtools (Li *et al.* 2009) were used to sort and index the aligned bam files, and finally bedtools (Quinlan and Hall 2010) were used to generate the coverage depth statistics. Plotting of the depth of coverage was done in R (Pereira *et al.* 2017) in the RStudio (RStudio Team 2016).

Global comparison dataset

All complete and partial genome sequences available in GenBank Nucleotide database (https://www.ncbi.nlm.nih.gov/genbank/) as on 19/09/2017 were used to prepare a global RSV-A genotype ON1 genomic and G-gene dataset. To prepare the global ON1 dataset, all RSV sequences from GenBank (search terms: respiratory syncytial virus) were downloaded, created a local blast database in Geneious (Eiter *et al.* 2003), and performed a local blast search using the 144 nucleotide sequence region of the ON1 genotype. To remove duplicates, the sequences were binned by country of sample collection, filtered of duplicates and then re-collated into a single dataset. For the global G-gene dataset of 1,167 sequences, the sequence length ranged from 238-690bp. The final alignment of 344 ON1 genome sequences comprised the sequences reported in this study (n=154) and additional publicly available GenBank ON1 sequences (n=190). In addition to the ON1 genomes, a total of 30 genotype GA2 genome sequences from Kilifi were generated. The alignments were inspected in AliView (Larsson 2014) and edited manually removing unexpected spurious frame-shift indels (largely homopolymeric and most likely sequencing errors).

Maximum likelihood phylogenetic analyses and root-to-tip regression

Separate Maximum-Likelihood (ML) phylogenetic trees were generated using multiple sequence alignments of the three datasets, i.e. Kilifi WGS, and global G-gene and WGS datasets. The ML trees were inferred using both PhyML and RaxML, with each optimizing various parts of the tree generation process (i.e. borrowing strengths of both approaches), using the script generated and deposited by Andrew Rambaut at (https://github.com/ebov/space-time/tree/master/Data/phyml_raxml_ML.sh). The GTR+G model was used after determination as the best substitution model by IQ-TREE v.1.4.2 (Chernomor, von Haeseler and Minh 2016).

To determine presence of temporal signal ('clockiness') in our datasets, TempEst v1.5 (Rambaut *et al.* 2016) was used to explore the relationship between root-to-tip divergence and sample dates. The data were exported to R (Pereira *et al.* 2017) to perform a regression with the 'lm' function.

Estimating the number of local variant introductions

To differentiate between local variants arising from a recent introduction and imported variants with greater genetic differences than is expected from local diversification, a pragmatic criterion previously described by *Agoti et al.* in (Agoti *et al.* 2015b) was used. Briefly, a variant is a virus (or a group of viruses) within a genotype that possesses $N_d \geq 10$ nucleotide difference compared to other viruses. This N_d nucleotide difference is a product of the length of the genomic region analyzed, estimated substitution rate for that region, and time. This analysis was done using usearch v8.1.1861 (Edgar 2010).

Protein substitution and selection analysis

Using the aligned Kilifi (ON1 and GA2) genome dataset, patterns of change in nucleotides (single nucleotide polymorphisms or SNPs) and amino acids were sought using Geneious v11.1.2 (Eiter *et al.* 2003) and BioEdit 7.2.5 (Hall 1999), respectively. Potential positively selected and co-evolving sites within the coding regions were identified using HyPhy (Pond, Frost and Muse 2005) and phyphy (J. Spielman 2018). SNPs were called from both the complete dataset and from an alignment of the consensus sequences from GA2 and ON1, whereby a consensus nucleotide was determined as the majority base at a given position. For the positive selection analysis, two strategies were used; gene-wide selection detection [BUSTED (Murrell *et al.* 2015)] and site-specific selection [SLAC, FEL (Kosakovsky Pond *et al.* 2005), FUBAR (Murrell *et al.* 2013) and MEME (Murrell *et al.* 2012)]. Codon positions with a p-value <0.1 for either the SLAC, FEL and MEME models or with a posterior of probability >0.9 for the FUBAR method were considered to be under positive selection.

Bayesian phylogenetics

To infer time-structured phylogenies, Bayesian phylogenetic analyses were performed using BEAST v.1.8.4 (Drummond *et al.* 2012). Because of sparse data at the 5' and 3' termini and in the non-coding regions of the genomic datasets, only the coding sequences (CDS) were used as input. The SRD06 substitution model (Shapiro, Rambaut and Drummond 2006) was used on the CDS and three coalescent tree priors were tested, i.e. a constant-size population, an exponential growth population, and a Bayesian Skyline (Drummond *et al.* 2005). For each of these tree priors, combinations with the strict clock model and an uncorrelated relaxed clock model with log-normal

distribution (UCLN) (Drummond *et al.* 2006) were tested with the molecular clock rate set to use a non-informative continuous time Markov chain rate reference prior (CTMC) (Ferreira and Suchard 2008). For each of the molecular clock and coalescent model combinations, the analyses were run for 150 million Markov Chain Monte Carlo (MCMC) steps and performed both path-sampling (PS) and stepping-stone (SS) to estimate marginal likelihood (Baele *et al.* 2012, 2013). The best fitting model was a relaxed clock with a Skyline coalescent model, *Table 4.1*.

Table 4.1: Marginal likelihood estimation of the best clock and coalescent models

Molecular Clock + Population Model	**Path-Sampling (PS)**	**Stepping-Stone (SS)**
Strict + Skyline	-40221.86	-40237
Relaxed + Skyline	**-40203.78***	**-40212.06***
Strict + Exponential	-40285.47	-40293.39
Strict + Constant	-40284.81	-40293.96
Relaxed + Constant	-40263.39	-40275.27
Relaxed + Exponential	-40272.2	-40283.01

*The highest marginal likelihood estimates for each analysis

BEAST was then run with 300-400 million MCMC steps using the SRD06 substitution model, Skyline tree prior, and relaxed clock model to estimate Bayesian phylogenies. For the time to the most recent common ancestor (TMRCA) estimates, the same substitution model and tree prior were used as above but with a strict clock model. For the global G-gene dataset, BEAST was run with 400 million MCMC steps using the HKY substitution model, Skyline tree prior, and a relaxed clock model. All the BEAST analyses were performed using the Broad-platform Evolutionary Analysis General Likelihood Evaluator (BEAGLE) library to enhance computation speed (Suchard and Rambaut 2009; Ayres *et al.* 2012). Tracer v1.6 was used to check for

convergence of MCMC chains and to summarize substitution rates. Maximum clade credibility (MCC) trees were identified using TreeAnnotator v1.8.4 after removal of 10% burn-in and then visualized in FigTree v1.4.3.

Principal component analysis

To check on any clustering and stratification patterns, principal component analysis (PCA) was performed using the R package FactoMineR (Lê, Josse and Husson 2008). The input data were a matrix of pairwise distances from genome sequence alignment using the "N" model of DNA evolution, i.e. the proportion or the number of sites that differ between each pair of sequences. Each genome on the PCA plot was annotated by the continent of sample origin.

4.4 Results

4.4.1 ***Genome sequencing and assemblies***

Over the five RSV epidemics sampled (2011/2012 to 2015/2016), a total of 3,157 samples were collected from eligible children at KCH, 3,146 (99.7%) were tested for RSV by IFAT or real-time PCR, and 801 (25.5%) RSV positives identified. Of these, 434 (54.2%) were RSV-A, of which 412 (94.9%) were successfully sequenced from routine G-gene sequencing, with 354 (85.9%) of genotype ON1 and the remainder 58 of genotype GA2. From the peripheral health centres within the KHDSS, a total of 32 RSV-A positives were identified by real-time PCR.

A total of 184 RSV-A genomes were generated in this study, comprising genotypes ON1 (*n*=154) and GA2 (*n*=30); *Appendix 7.5*. This dataset included 176 genomes from inpatients at KCH and 8 genomes from the KHDSS. The sequencing success for

KCH samples was 87% (154 full genomes /177 samples processed for sequencing) for ON1 viruses (the denominator a 50% sub-sample of all 354) and 52% (30 full genomes/58 samples processed for sequencing) for GA2 viruses, and for KHDSS samples was 25% (8 full genomes /32 samples processed for sequencing). The Ct values for KHDSS samples (as an indicator of viral load) had a median of 26.3 (IQR: 22.9 - 28.0), which was slightly higher than for the KCH sample set with a median Ct of 24.4 (IQR: 22.2 - 26.9), *Figure 4.1*. The differential sequencing success rate between the KCH and KHDSS samples may have been due to the slightly lower viral loads in the KHDSS samples. Between 0.2 to 4.3 million short reads were available per sample of which RSV specific reads ranged between 0.001 to 3.9 million reads (each of 250 bases). The genome assemblies had a median length of 15,054 nucleotides (range: 13,966-15,322) and mean depth of base coverage per genome ranging from 39 to 66,457 (calculated from e.g. [3.9m reads X 250 bases]/15,000).

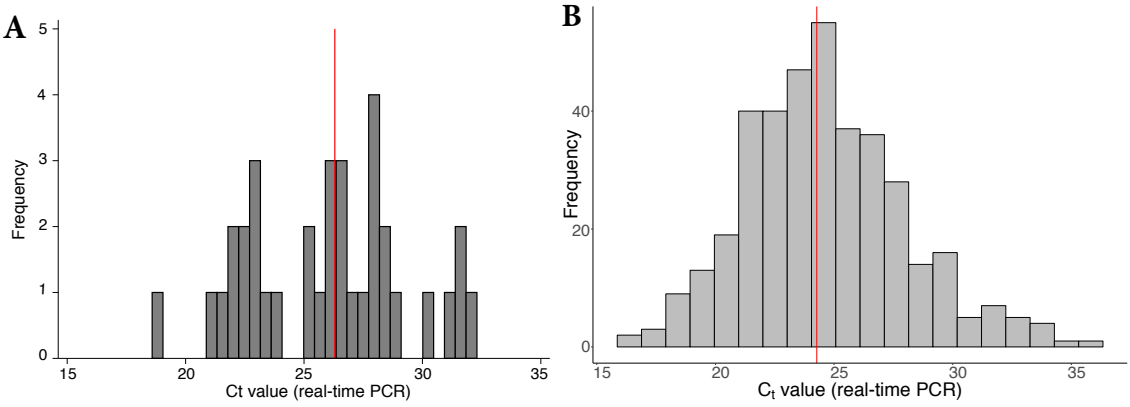


Figure 4.1: Histograms showing distribution of PCR Ct values (red line = median) for samples collected in Kilifi between 2011 and 2016 from the (A) KHDSS and at the (B) KCH.

Whereas the samples targeted for WGS were generally of high viral content (lower Ct value), it is apparent there was reduced genome yield (proportion of genome assembled) from samples with lower viral loads (i.e. higher Ct values); *Fig 4.2A*. However, the samples successfully sequenced and analyzed here generally had lower Ct values (higher viral loads) as shown in *Fig 4.2B*. The median fraction of the genome with unambiguous base calls was 98% with reference length from KC731482. Read coverage across the genomes was non-uniform, *Fig 4.2C,* suggesting varied PCR amplification efficiency among primer pair combinations combined with increased sequencing yield from the ends of the amplicons.

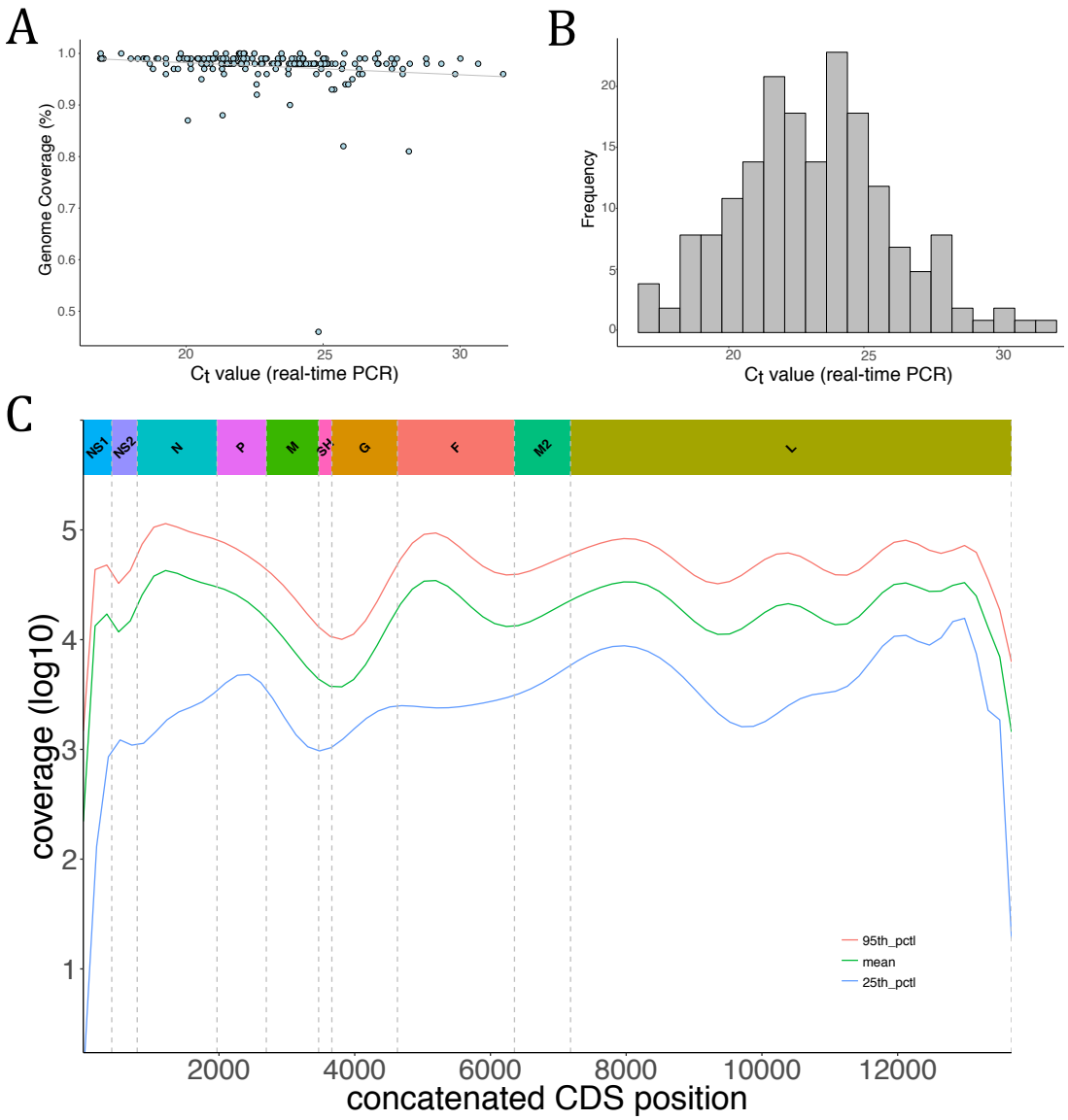


Figure 4.2: Kilifi RSV-A 2012-2016 sequenced genomes fraction, Ct value and coverage (A) The proportion of RSV genome length sequence recovered (using KC731482 as the reference) for all the 184 genomes was plotted as a function of sample's diagnostic real-time PCR Ct value. (B) The distribution of the diagnostic real-time PCR Ct values for the 184 sequenced samples reported here (KCH and KHDSS). (C) Shows the log (base 10) values of the sequencing depth (see Methods) at each position of the genome assemblies along the concatenated RSV ORFs (i.e. excluding the intergenic regions).

4.4.2 ***Bayesian reconstruction of ON1 epidemiological and evolutionary history***

The global ON1 whole-genome MCC phylogenetic tree, *Figure 4.3A*, shows evolutionary relationship among ON1 viruses from five sampled continents. The TMRCA of the ON1 strains from the most recent tip (7 April 2016) was estimated to be 11.07 years [95% HPD: 9.85-12.31], resulting in an estimated ON1 emergence date of between December 2003 and June 2006. This estimated date of emergence is earlier than a previous estimate (2008-2009) using the G-gene alone (Duvvuri *et al.* 2015), but such a difference could be a reflection of the different datasets (by geography and sampling time frame). *Comas-Garcia et al.* have reported the earliest ON1 strain identified to date in November 2009 from central Mexico (Comas-García *et al.* 2018), and from our estimates this suggests a period of 3-6 years of circulation of this virus before first detection. The genome-wide substitution rate for the ON1 viruses was estimated at 5.97 X 10^{-4} nucleotide substitutions per site per year [95% HPD: 5.42-6.58 X 10^{-4}], similar to previous estimates for RSV group A full length sequences sampled over several epidemics (Tan *et al.* 2013; Agoti *et al.* 2015a) but slower than estimates from whole-genomes sequences of samples collected from a household study over a single RSV epidemic (2009-2010) within the same Kilifi location [2.307×10^{-3}] (Agoti *et al.* 2017) and from using a global ON1 G-gene dataset [4.10×10^{-3}] (Duvvuri *et al.* 2015). Across the genome, estimates of evolutionary rates for individual ON1 open reading frames (ORFs) varied, *Figure 4.3B*, with the mean substitution rate highest in the G-gene, lowest in NS1, and moderate (with tight 95% HPD intervals) for the whole genome.

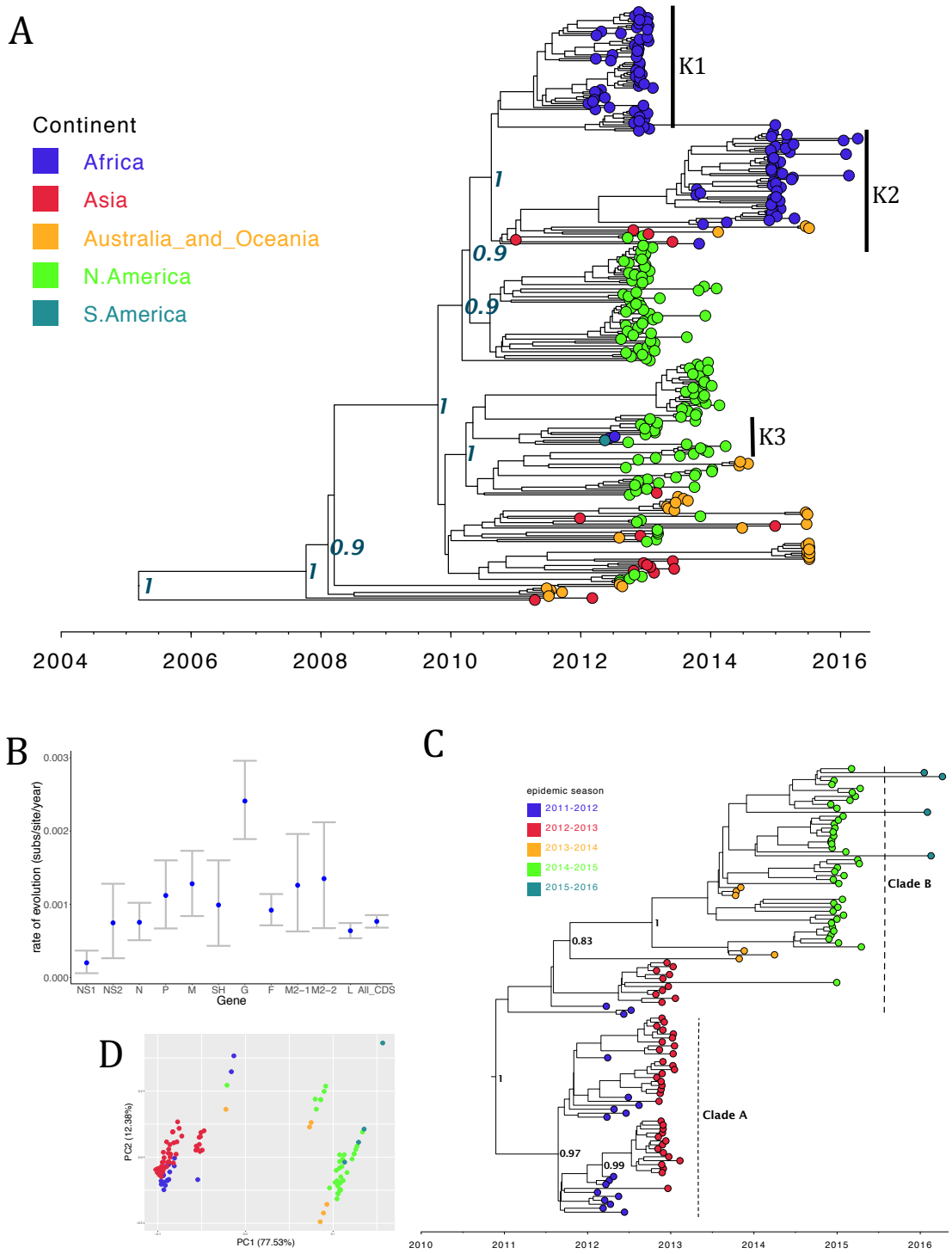


Figure 4.3: WGS MCC trees and PCA showing global and local clustering of ON1 viruses (*A*) Maximum clade credibility tree inferred from 344 global full genome sequences (see Methods) with the tips colour coded with the continent of sample collection. All the African samples (in blue, K1-3 and vertical bars) in this dataset were only available from Kilifi (Kenya). Node labels are posterior probabilities indicating support for the selected nodes. (*B*) shows the evolutionary rate estimates for the different genotype ON1 ORFs. (*C*) is an MCC tree inferred from 154 ON1 genomes from Kilifi annotated with identified lineages A and B, and the tips colour coded with the epidemic season. (*D*) is a PCA analysis (see Methods) of

the same dataset as (*C*) and similarly annotated with the epidemic season. Percentage of variance explained by each component is indicated on the axis.

The Kilifi ON1 genomes were placed into three lineages (K1-3 and black vertical bars) on the global tree in *Figure 4.3A*. However, when the Kilifi ON1 WGS were analyzed separately, *Figure 4.3C*, two lineages were observed (labelled A and B) with a temporal grouping whereby A comprised sequences from the 2011-2013 RSV epidemic period while B comprised sequences predominantly from the epidemic period 2013-2016. These lineages and temporal patterns are further highlighted by the PCA analysis in *Figure 4.3D*. Based on the phylogenetic placement of the Kilifi ON1 genomes on the global tree in *Figure 4.3A*, it was estimated that there could have been at least three separate introductions of ON1 viruses into Kilifi. One of these potential Kilifi ON1 introductions (K3) was characterized by only two cases, which is consistent with limited local transmission. In addition, the eight outpatient ON1 viruses collected from the KHDSS were interspersed with viruses sampled from inpatient admissions at KCH suggesting that our sampling at the hospital might be representative of the larger KHDSS community.

Using the global whole genome ON1 substitution rate estimate above, the Kilifi ON1 genomes dataset (length 15,404 bp) and a pragmatic criterion previously described by *Agoti et al.* in (Agoti *et al.* 2015b) to differentiate between local and imported variants, it was estimated that there could have been up to 73 ON1 introductions into Kilifi implying that lineages K1 and K2 in *Figure 4.3A* are not just two initial introductions that have grown over time but comprise multiple introductions. Even when the higher substitution rate previously estimated from ON1 partial G-gene sequences by *Duvvuri et al* in (Duvvuri *et al.* 2015) was used, i.e. 4.10×10^{-3}

substitutions/site/year which translates to a difference of at least 63 nucleotides between any two genomes to be classified as separate introductions, this resulted in an estimate of 6 separate introductions. This suggests that multiple seeding introductions of ON1 viruses may have been required to sustain their local transmission.

4.4.3 ***Placement of Kilifi ON1 viruses in the global context using G gene***

As there are far more partial G gene sequences than full genomes, we explored the placement of Kilifi ON1 viruses in the global context using a set of 1,167 global G gene sequences. The global G gene MCC tree is shown in *Figure 4.4A* with the corresponding sampling locations in *Figure 4.4B*. Similar to *Figure 4.3A*, the Kilifi viruses in *Fig 4.4A* were placed in multiple (perhaps 4) lineages further supporting the idea of multiple introductions into Kilifi. On the contrary, viruses from each of the other two African countries represented (Nigeria and South Africa) were restricted to single major branches even though this could be as a result of the very few ON1 sequences available from these countries (<10 from each). Furthermore, viruses closely related to the ON1 viruses with limited local transmission in Kilifi described above were frequently isolated in other locations. With the ON1 viruses assigned the corresponding continent of sample collection (*Figure 4.4A*), there was neither a single major branch on the tree comprised solely of viruses from a specific continent nor a continent whose viruses were only found within a single major branch, suggesting both intra and inter-continental circulation patterns. However, the majority of the Kilifi ON1 viruses in *Figure 4.4A* clustered with European viruses with a few others clustering with Asian viruses suggesting perhaps a predominantly European source of RSV introductions into Kilifi.

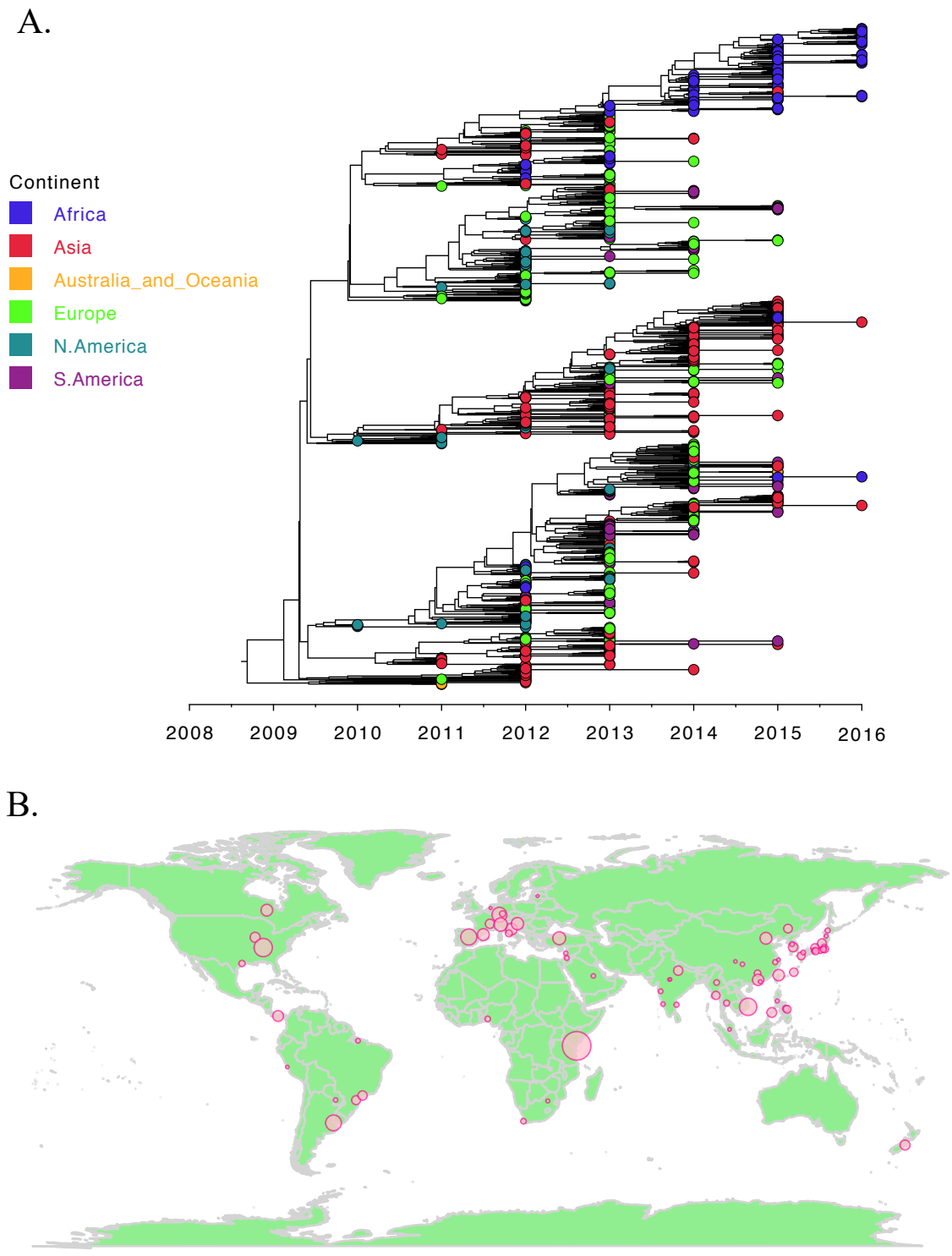


Figure 4.4: Relative sampling and placement of Kilifi ON1 viruses on a global MCC tree. (*A*) An MCC tree inferred from 1,167 partial ON1 G gene sequences with the tips colour coded with the source continent. (*B*) Sampling locations of the dataset in (*A*) with circles representative of relative proportion of contributing sequences by country.

4.4.4 ***Genomic diversity of Kilifi RSV-A viruses***

Pairwise intra-genotypic genetic diversity analysis of the GA2 and ON1 genomes from Kilifi, *Figure 4.5A* and *4.5B*, show unimodal and bimodal distributions respectively consistent with two genetically distinct circulating strains of ON1 viruses. Analyzing for substitutions across the genomes by entropy plots (*Figure 4.5C*), a total of 746 single nucleotide polymorphisms (SNPs) were identified with frequencies of >1% in the set of 184 genomes. Of these SNPs, the majority (589, 78.9%) were found within coding sequences/regions (CDS). The three CDSs with the most substitutions were the polymerase L (39.6%), the glycoprotein G (14.8%) and the fusion F protein (14.6%). However, when the SNP count is considered against CDS length then G, SH and M2-2 top the first three slots. Only 145/589 (24.6%) of these coding mutations resulted in non-synonymous changes, *Appendix 7.6*. The majority of the non-synonymous mutations occurred within the G, SH and M2-2.

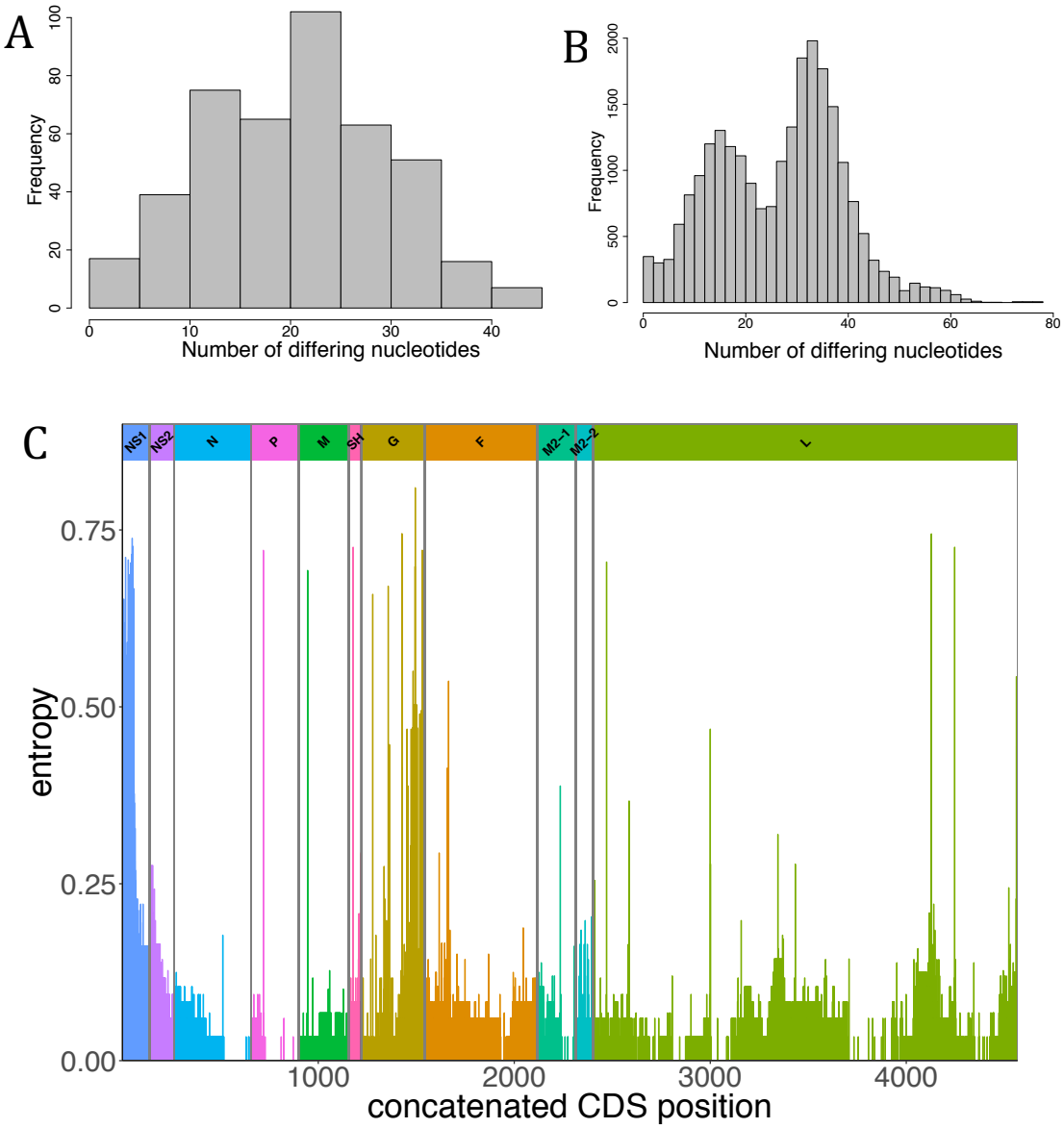


Figure 4.5: Pairwise genomic distances and genome-wide amino acid variation

The distribution of pairwise genetic distances between genotype GA2 and ON1 genome sequences are shown in (*A*) and (*B*), respectively. (*C*) is an entropy plot showing amino acid variation along the concatenated ORFs of Kilifi RSV-A genomes.

4.4.5 ***Phylogenetic divergence between ON1 and GA2 viruses***

The currently known or *de facto* distinguishing feature of the ON1 from GA2 strains is the 72-nucleotide duplication within the G gene. It has been shown from phylogenetic analysis of the G-gene that RSV-A genotypes form distinct clusters (Peret *et al.* 1998). However, it has not been investigated if the distinct clustering is replicated in the other genes especially for the closely related genotypes GA2 and ON1 viruses. An exploratory root-to-tip regression analysis of ORF specific ML trees, whose topologies were similar to the MCC BEAST trees described herein, confirmed that all but the NS1, NS2 and SH proteins had good temporal signals, *Figure 4.6*.

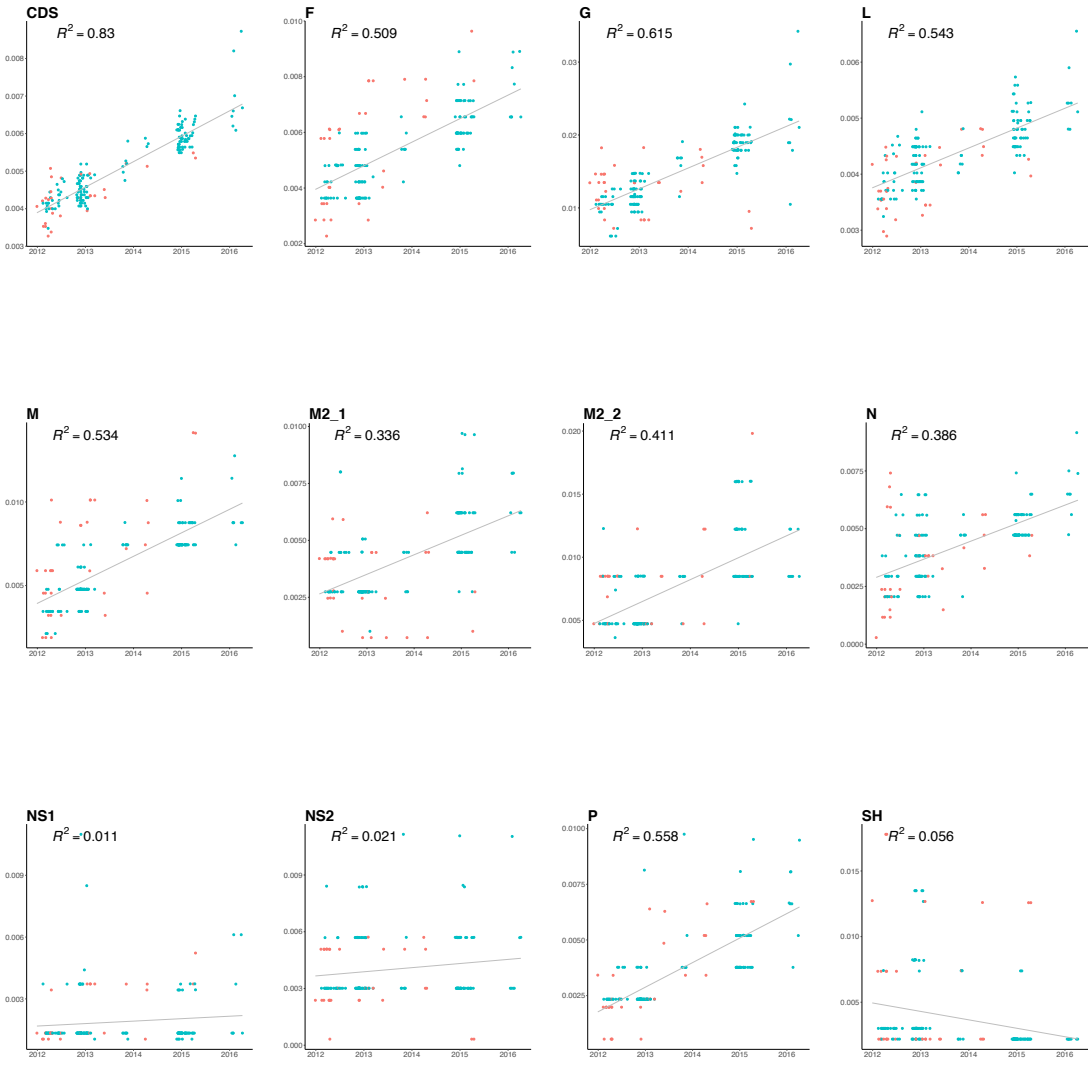


Figure 4.6: Root-to-tip regression analysis of Kilifi RSV-A genomes ORFs.
GA2 and ON1 sequences are shown by red and cyan dots, respectively.

To assess if the 72-nucleotide duplication is the only marker of the ON1 strains or a complementary mutation, the 11 RSV ORFs individually and a concatenated set of 10 ORFs (excluding the G) were analyzed. Distinct and well supported ON1 and GA2 clusters were observed in the concatenated set of 10 ORFs as well as in five individual coding regions (F, G, L, N and P), *Appendix 7.7*, confirming that genetic markers outside G also differentiate the ON1 and GA2 genotypes. The node posterior support, however, for divergence between GA2 and ON1 was quite low (50-70%) in the N and P proteins despite observation of distinct clusters. Nonetheless, determining the order in which the GA2-ON1 divergence in the five ORFs might have occurred was not feasible from this analysis as the divergence could have occurred anywhere on the branch between the GA2-ON1 split time and ON1 TMRCA in *Figure 4.7A*. This divergence chronology dilemma is highlighted by the overlapping MRCA estimates for the individual ORFs of between 2007-2011 [95% HPD: 2004.59-2012.06] in *Figure 4.7B*.

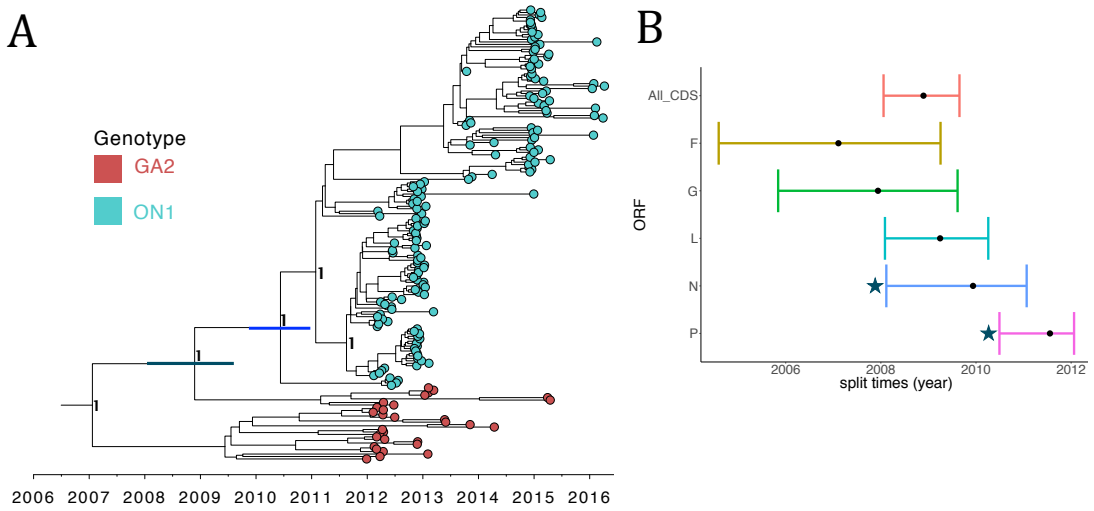


Figure 4.7: Estimated TMRCA for Kilifi RSV-A viruses and ORFs

(*A*) MCC tree inferred from 184 RSV-A complete genome sequences (concatenated coding regions only) from Kilifi with the tips colour coded by genotype, i.e. ON1 (cyan) and GA2 (red). The two node bars indicate the 95% HPD interval for the TMRCA for the Kilifi GA2 and ON1 viruses (grey), and Kilifi ON1 strains (blue). Node labels are posterior probabilities indicating support for the selected nodes. (*B*) shows the TMRCA (with 95% HPD interval) of the node separating Kilifi RSV-A genotype GA2 and ON1 viruses for a concatenated set of all ORFs (red) and five different ORFs (F: orange, G: green, l: cyan, N: blue, and P: purple). The stars (*) indicate node posterior support of less than 0.9 (i.e. low support) for the split between GA2 and ON1 in the nucleoprotein (N) and phosphoprotein (P) ORFs.

4.4.6 ***Signature substitutions distinguishing ON1 from GA2 viruses***

Through a comparative genome-wide scan along the RSV-A coding genome, we analyzed for SNPs between the consensus Kilifi ON1 and GA2 viruses. A total of 66 signature nucleotide substitutions (defined as SNPs differentiating ON1 from GA2 viruses) were identified, highlighted in yellow in *Appendix 7.6*. While the majority of these signature substitutions were synonymous, 14 were non-synonymous substitutions (*Table 4.2*); nine in the G protein, two each in the F and L proteins, and one in the M2-1 protein. However, these signature substitutions had no effect on our RSV multiplex real-time PCR diagnostics as they occur outside the target primer binding sites in the N gene. Changes at the codon sites 142 and 237 of the G protein

have previously been shown to characterize antibody escape mutants, and were located within strain-specific epitopes (Martínez, Dopazo and Melero 1997). The two signature substitutions in the F protein (116 and 122) occur within site p27, which is the most variable antigenic site in the F protein (Hause *et al.* 2017).

Table 4.2: Signature non-synonymous substitutions between genotype ON1 and GA2 viruses

ORF	[α]ORF Nt Pos.	ORF AA Pos.	Change	AA Change	SNP Type
G	424	142	TT -> CA	L -> Q	Substitution
G	622	208	C -> A	L -> I	Transversion
G	695	232	G -> A	G -> E	Transition
G	709	237	A -> G	N -> D	Transition
G	758	253	A -> C	K -> T	Transversion
G	817	**273[β]**	T -> A	Y -> N	Transversion
G	821	274	C -> T	P -> L	Transition
G	851	284	72 nt duplication	24 AA insertion	Deletion
G	929 (GA2: 857)	**310[β]**	C -> T	P -> L	Transition
G	941 (GA2: 869)	314	T -> C	L -> P	Transition
F	346	116	A -> G	N -> D	Transition
F	364	122	G -> A	A -> T	Transition
M2-1	349	117	A -> C	N -> H	Transversion
L	1792	598	C -> T	H -> Y	Transition
L	5175	1725	A -> T	E -> D	Transversion

ORF=Open Reading Frame, Nt=Nucleotide, AA=Amino Acid, Pos.=Position

[α]Positions are relative to ON1 strains, in which complementary positions in GA2 (without the duplication) within the G protein are shown in brackets.

[β]Positively selected sites

4.4.7 ***Signature substitutions between ON1 lineages with successful and limited local transmission***

A similar genome-wide comparative scan between the consensus of genomes of viruses with successful (K1 and K2) and those with limited local transmission (K3) was performed for characteristic signature polymorphisms. A total of 33 SNPs were identified between these two groups of viruses, *Appendix 7.8*, of which nine resulted in non-synonymous changes; five in G, two in F and one each in M2-2 and L. In three of these nine non-synonymous SNPs, the K3 viruses shared substitutions with the GA2 viruses (G: codons P274L and P310L, and F: codon A122T). Whether these polymorphisms are neutral mutations or influence local transmission of the virus warrants further investigation.

4.4.8 ***Signature substitutions distinguishing BA from non-BA viruses***

Similar to the previous two analyses above, we analyzed for SNPs between consensus global BA and non-BA viruses. A total of 39 signature non-synonymous substitutions were identified across nine RSV proteins, i.e. NS1, N, M, SH, G, F, M2-1, M2-2, and L, with most of the substitutions in the G (15 substitutions) and L (13 substitutions), *Appendix 7.9*.

4.4.9 ***Nature of ON1 emergence: Multiple or single duplication event?***

Two papers studying the molecular evolution of RSV genotype ON1 have concluded that the emergence of ON1 happened multiple times or in a convergent manner (Schobel *et al.* 2016; Comas-García *et al.* 2018). These conclusions were based on the observation of clustering of some ON1 sequences amongst non-ON1 sequences. In addition, I have had discussions with colleagues working on RSV from Argentina

who had noticed some of their GA2 viruses phylogenetically placed with ON1 viruses and concluded that this might imply that these ON1 viruses had lost the 72nt duplication. None of these observations were made with the Kilifi dataset. However, based on these reports and discussions, an analysis of the global RSV sequence datasets was undertaken to investigate these conclusions.

Similar to observations by Schobel *et al.* and Comas-Garcia *et al.*, some of the ON1 viruses in this analysis were phylogenetically placed within the non-ON1 cluster both for the WGS and G-gene global datasets, *Figure 4.8*. Surprisingly, there was a trend in the sources of these oddly placed ON1 viruses. For the WGS dataset, even though collected from Jordan and New Zealand between 2011 and 2012, they were sequenced at the J. Craig Venter Institute (JCVI) using Illumina sequencing technology, assembled using a referenced-based assembly (clc_ref_assemble_long v. 3.22.5507), and did not form a single cluster within the GA2 viruses, *Figure 4.8*. With the global G-gene dataset, the G-gene regions extracted from WGS of ON1 viruses placed within the GA2 cluster described above still clustered with GA2 viruses and also did not form a single cluster, *Figure 4.8*. In addition to these, there were two ON1 sequences each from Argentina (Rojo *et al.* 2017) and Spain (unpublished) collected in 2014 and 2015 that were placed with GA2 viruses with the four sequences forming a single cluster. However, there were no Spanish or Argentinian sequences available post 2015 that would illuminate on whether this cluster has grown over time.

We also looked at the non-ON1 viruses placed amongst ON1 viruses based on WGS assemblies. For the WGS dataset, they were collected in the US and came from JCVI,

Broad MIT (Illumina sequencing and assembled with Vicuna) and University of Washington (assembly method not stated). They did not form a single cluster but were interspersed within the ON1 viruses. For the G-gene dataset, the non-ON1 viruses clustering with ON1 viruses included those extracted from the non-WGS sequences above and one sequence each from Brazil (2011, University of Sao Paulo) and Taiwan (2015), and these too did not form a single cluster.

We hypothesized that since the RSV-B genotype BA 60nt duplication emerged first, we might observe the phenomenon of multiple independent emergence of the 60nt duplication and/or loss of the duplication in the RSV-B sequence dataset as well. For the WGS dataset, *Figure 4.9*, there were five BA viruses clustering with non-BA viruses from the JCVI collected in the US between 1996 and 1998. There were also two sequences from the JCVI (Jordan 2010 and USA 2013, not in a single cluster) and one sequence from Brazil (2010, Illumina, Spades, University of Sao Paulo) that were non-BA but clustered with the BA viruses. For the G-gene dataset, one of the BA viruses in non-BA cluster came from Brazil (collected in 2003, University of Sao Paulo) with the other five sequences from the JCVI and collected between 1996 and 1998. For the non-BA strains clustering with BA viruses in this G dataset, they included G-gene sequences extracted from the three WGS described above and another two sequences from Kilifi Kenya (in a single cluster and previously suspected to have lost the 60nt duplication (Agoti *et al.* 2010)).

Finally, as distinct clustering between ON1 and GA2 viruses had been observed in five individual coding regions (F, G, L, N and P) within the Kilifi RSV-A dataset we analyzed the global RSV-A WGS dataset. Similar observations were made with the

global dataset as with the Kilifi dataset, *Appendix 7.10*, with distinct clusters between ON1 and non-ON1 viruses in the five coding regions save for the sequences from the JCVI, Broad MIT and University of Washington as described above whose placements were out of order with their respective genotypes.

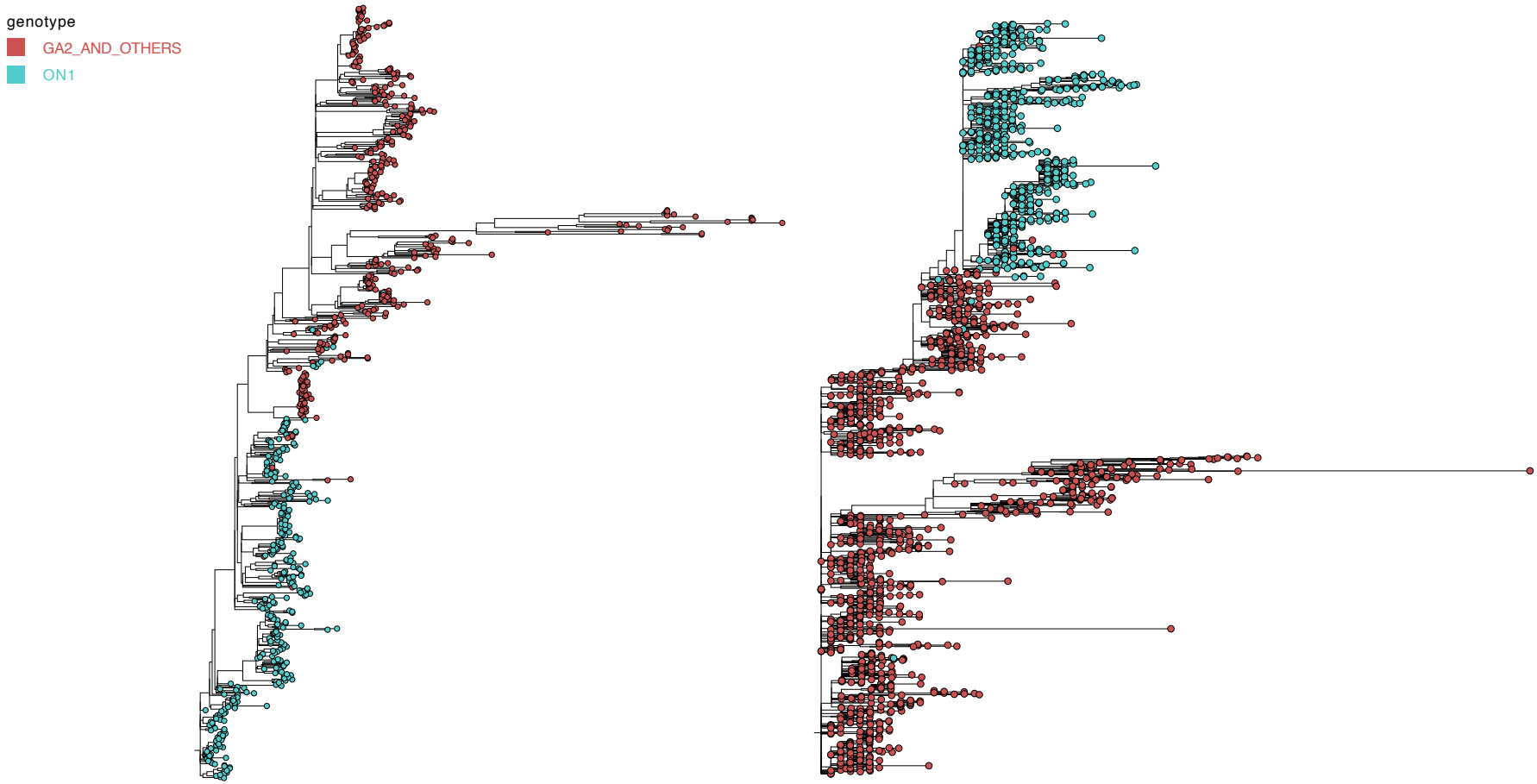


Figure 4.8: Global RSV-A WGS and G-gene ML trees showing phylogenetic clustering between ON1 and other RSV-A genotypes

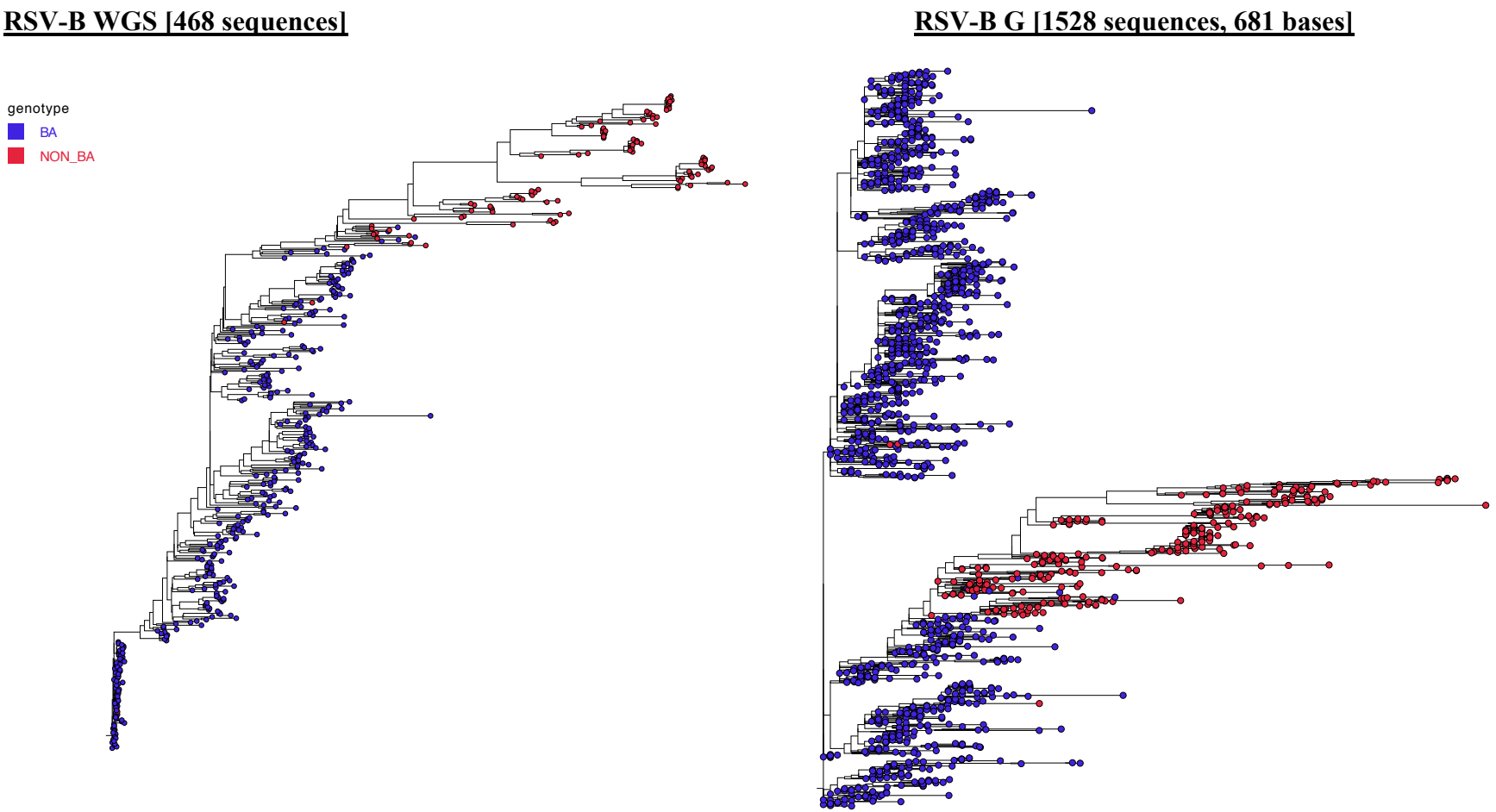


Figure 4.9: Global RSV-B WGS and G-gene ML trees showing phylogenetic clustering between BA and other RSV-B genotypes

4.4.10 ***Patterns of selective pressure across the RSV-A genomes***

We conducted selection analysis on all 11 RSV ORFs for the dataset, *Appendix 7.11*. ORF-wide episodic diversifying selection was only detected in the NS1 and M proteins. A total of nine positively selected codon sites were identified within the G (73, 201, 250, 251, 273, 310), NS2 (15) and the L (2030, 2122) by at least one method, with site 310 in the G identified as positively selected by all the four methods. Notably, sites 273 and 310 (shown in bold in *Table 4.2*) within the G protein detected to be under positive selection were also identified as signature SNPs. However, the number of positively selected sites could have been underestimated in the analysis that was limited to Kilifi RSV-A genomes and care should be taken while interpreting these results as some of the positively selected sites were only detected by one method and at default (less stringent) cut-offs.

4.5 Discussion and Conclusions

In this chapter, we report an in-depth analysis of local and global RSV genotype ON1 evolution and transmission using whole genome sequence data. We describe RSV-A genomic diversity and identify polymorphisms with the most potential in influencing RSV evolution and phenotype. Utilizing genomes from samples collected between 2010–2016, including 184 complete genomes from Kilifi alone, we obtained a finer resolution on the pattern of RSV introductions, persistence and evolution in a defined location, and the changes within the genome that might be important for the persistent circulation of the virus.

Genetic variation not only provides important insights into RSV relatedness by which to infer transmission events but also highlights potential functional changes in the

virus. From this analysis, it was shown that substitutions are widespread across the RSV genome (both for RSV-A and B) but occur at higher frequency within the structural proteins (G and F) and in the polymerase (L). The G protein has the most genetic flexibility of the RSV ORFs to accommodate frequent substitutions including large duplications, and previous studies have described epitope positions associated with escape using specific monoclonal antibodies or in natural isolates (García *et al.* 1994; Cane and Pringle 1995; Cane 1997; Martínez, Dopazo and Melero 1997). The F protein site p27 with the two signature GA2-ON1 substitutions has been shown to (i) be involved in the host cell entry of RSV that involves micropinocytosis, followed by proteolytic activation of the F protein at the second furin-cleavage site and release of the p27 peptide (Krzyzaniak *et al.* 2013), and (ii) possess greater binding affinity for serum antibodies from young children (<2 years) than any of the other antigenic sites in the F protein and may be responsible for group specific immunity that distinguish between RSV-A and RSV-B viruses (Fuentes *et al.* 2016). The implications of observed substitutions in the L protein of the ON1 viruses remain unclear. However, considering both its role in genome replication and the emergence of the 72-nucleotide duplication in the G ORF, it is hereby proposed that either (i) these polymorphisms might have resulted in a sloppy polymerase, that resulted in a slip that generated the 72-nucleotide duplication in the G ORF (Komissarova and Kashlev 1997), or (ii) the 72-nucleotide duplication in the G may present a metabolic challenge for replicating a large genome and thereby facilitate adaptive polymorphisms within the polymerase (Canchaya *et al.* 2003). While a considerable number of SNPs were also found in ORFs other than the G, F and L proteins, only a very minor proportion of those changes resulted in amino acid substitutions implying very strong purifying selection.

Based on distinct phylogenetic clustering of ON1 and GA2 viruses in five ORFs within the Kilifi and global dataset, the emergence of ON1 is certainly characterized by additional substitutions across the genome in addition to the 72-nucleotide duplication within the G gene. And the same can be said of the emergence of the BA genotype. Assuming ON1 diverged from GA2 and through a single ancestral virus, it is unclear whether the multiple signature substitutions differentiating ON1 from GA2 viruses all arose from that single split event or have been acquired progressively over time. In case of the latter, the chronology of changes across the different ORFs is unclear. Understanding how and which mutations define the emergence of a new RSV variant may be important in describing substitutions that are either crucial for the survival of the variant and/or of some complementary structural or functional integrity. It is also likely that some of these substitutions are nothing more than genetic hitchhikers. Notwithstanding this lack of clarity on ON1 emergence, it has been shown for influenza A viruses that linked selection amongst antigenic and non-antigenic genes influences the evolutionary dynamics of novel antigenic variants (Raghwani, Thompson and Koelle 2017). Further, it has been demonstrated experimentally that adaptive evolution is a multi-step process that occurs in waves (Stern *et al.* 2017). The initial adaptive wave is thought to occur rapidly and is characterized by founder or gatekeeper mutations. Thereafter, additional waves of evolutionary fine-tuning occur (Grubaugh and Andersen 2017). Similar studies in RSV would be important to determine if such dynamics do characterize RSV's evolutionary history and may also inform the design of an RSV vaccine.

From recent publications and personal discussions, it is not clear if the emergence of ON1 occurred once through a single ancestral ON1 virus or multiple times (Schobel

et al. 2016; Comas-García *et al.* 2018). From our analysis using global RSV-A and RSV-B datasets, we find oddly placed sequences (by genotype) but whose sources were often the same labs and/or part of large sequencing projects. It is also interesting that these 'misplaced' viruses do not have descendants (except in the one case where four sequences clustered together). There are three likely reasons for these observations; (i) recombination, (ii) cross-talk (contamination) between samples during pre-processing or sequencing, and (iii) miss-assemblies. Recombination in RSV is unprecedented in natural isolates and has only very rarely been detected under controlled laboratory conditions (Spann, Collins and Teng 2003). A recombination analysis by *Tan et al* (Tan *et al.* 2012) showed that genomes in which recombination was detected, while sampled from natural infection, came from the same labs (Kumaria *et al.* 2011; Rebuffo-Scheer *et al.* 2011) and were likely to be either PCR or sequence assembly artefacts. Therefore, it is highly likely that the emergence of these duplication variants occurred only once in their evolutionary history. Further, since the discovery of RSV only the two large duplications, one of 60 nucleotides (BA) and the other of 72 nucleotides (ON1) have been detected. We think that if the phenomenon of multiple emergence of these duplications were likely and frequent in RSV, the same would have been evidenced in the emergence of the other RSV genotypes (e.g. some other group A genotypes such as GA2 also phylogenetically placed within GA5, etc) or even frequent detection of duplication variants. This analysis also highlights a potential problem from using sequences from large sequencing projects in analyses where the accuracy of the assemblies may be questionable.

ON1 is rapidly replacing GA2 in Kilifi, suggesting that this variant may have some fitness advantage in this location. We have however previously showed that genotype ON1 viruses did not result in higher risk of severe disease compared to GA2 viruses in Kilifi (Otieno *et al.* 2017). Globally, ON1 prevalence varies by location and there are conflicting reports with regards to differences in virulence between ON1 and GA2 strains (Panayiotou *et al.* 2014; Yoshihara *et al.* 2016). Even with the discordant results, which may be due to differences in study populations and analysis methods, there might be phenotypic differences between viruses belonging to these two genotypes. Identification of such phenotypic differences and the potential drivers might augment our current understanding of the pathogenesis of this virus. Expanded RSV surveillance in additional locations will offer better insight into the nature of these replacement dynamics.

Observations from this study using whole genomes reinforce previous findings based on partial G-gene sequences (Agoti *et al.* 2014b, 2015b; Otieno *et al.* 2016, 2017) that RSV epidemics are characterized by the introduction and circulation of multiple variants. In addition, persistence within the community seems to be sustained by only a proportion of these introductions. We have characterized genomic substitutions that distinguish between successful and dead-end ON1 introductions in Kilifi and find that the dead-end ON1 introductions share substitutions with the fast fading Kilifi GA2 strains. Nonetheless, it is evident that besides viral genetic factors there could be other determinants of successful onward transmission of a virus lineage. ON1 strains that were non-persistent in Kilifi were abundant in other parts of the world albeit with varied frequencies relative to other genotypes. Such determinants could include the host factors (e.g. births, immunity, genetics, contact patterns and mobility) and

environmental factors (e.g. temperature, rainfall and humidity) which warrant further investigations.

We live in times of rapid global movement of people, which may influence the spread of infectious diseases. The observation that most of the Kilifi sequences clustered with sequences from Europe and Asia suggests that RSV introductions into Kilifi originate predominantly from these two continents. It might not be surprising that Europe could be a source of RSV introduction into Kilifi, or a destination for viruses from Kilifi, considering that it accounts for the largest single group of tourists to Kenya (The Report: Kenya 2017 2017). In addition, the increasing Chinese economic interests in Africa (including Kenya) has resulted in an influx of Chinese into Africa for trade, work and tourism (The Economist 2013) and may account for the Asia-like ON1 strains. However, there are far too few partial ON1 sequences from Africa (only from Kenya, South Africa and Nigeria) and no ON1 genomes from outside Kilifi Kenya to help define intra-African transmission dynamics in detail. In fact, a recent study suggests that domestic tourism accounts for more than half of the growth in Kenya's tourism (Sunday 2018). As such, availability of sequences from across the country would be critical in deciphering if and how such tourist activities influence virus transmission patterns in Kenya. Such studies could be helpful in the design of future RSV transmission intervention strategies.

CHAPTER FIVE

5 Local and global RSV transmission dynamics

5.1 Introduction

From the preceding two chapters, it is apparent that RSV genotype ON1 transmission is characterized by frequent introductions and circulation of multiple variants within a community (Otieno *et al.* 2016, 2017, 2018). In addition, it was observed that most of the Kilifi sequences clustered with sequences from Europe and Asia which may suggest frequent exchange of viruses between Kilifi and these two continents. However, the origin of the viruses seeding the recurring RSV epidemics at varied geographic scales (e.g. community [Kilifi], local [Kenya] or continental [Africa]) and the factors that determine spread remain unclear. There has only been a single phylogeographic analysis of RSV spread in 2012 using partial G gene sequences (Katzov-Eckert *et al.* 2012). With the availability of more sequences (including whole genomes) and robust Bayesian phylogeographic methods, there was sufficient motivation to conduct a detailed analysis of the local and global patterns of the spatial spread of RSV.

In this chapter, RSV sequence data collected from across Kenya and other 50 global countries between 1977 and 2016 (inclusive) was used to estimate transition rates between discrete locations in a Bayesian statistical framework. To protect against sampling heterogeneity or bias in the sequence datasets, the transition rates were parameterized according to potential drivers of RSV spread (air travel or fluxes, depending on which fitted best) using a generalised linear model (GLM) within the same Bayesian framework (Lemey *et al.* 2009, 2014). This study highlights the

importance of integrating pathogen genetic and ecological information in improving understanding of the dynamics of infectious disease.

Analysis files and scripts can be found on GitHub: https://github.com/jrotieno/rsv_phylogeography/

5.2 Aims of the Chapter

We set out to determine the patterns of spread of RSV, locally within a country setting and between geographically defined regions (countries, continents and hemispheres).

5.3 Methods

Local (Kenyan) sample details and sequencing

We received a total of 400 RSV positive samples from CDC-K collected in Siaya County, Kakuma Refugee Camp and Dadaab Refugee Camp between January 2011 and June 2014 (see Chapter Two section 2.4.2 for an elaborate description of the study and sample details). While critical for this study due to the central location and its role as a major transport hub, samples from KNH (Nairobi) were not included in this analysis as they were received in Kilifi in November 2018 a time nearing thesis submission. However, these samples will be processed at a later date and sequences re-analyzed with the current dataset.

Nucleic acid extraction, G-gene amplification, confirmation of amplification success, and sequencing were performed as described in the Methods chapter (sections 2.7.2 and 2.7.3). However, to reduce costs and processing times, only the outer primers (AG20 and F164) were used for both G-gene amplification and sequencing reactions and thereby omitted the internal group specific nested PCR reactions. The sequencing

success of the two primers alone (see Results 5.4.1) was comparable to previous using four primers (see Chapter Three 3.4.1 and Chapter Four 4.4.1).

Global sequence dataset compilation

To prepare the global RSV datasets, all available RSV sequences from GenBank (https://www.ncbi.nlm.nih.gov/genbank/, search terms: respiratory syncytial virus) as on 22/05/2018 were retrieved and stored in Geneious (Eiter *et al.* 2003). All non-human RSV, lab strains, synthetic/vaccine related sequences, and sequences with no collection date and location were removed. To separate RSV-A from RSV-B sequences, a local blast search was performed using the 144 and 120 nucleotide sequence regions of the ON1 and BA genotypes, respectively. This method was better at separating sequences belonging to the two RSV groups compared to using group information annotation on different fields of GenBank files for separation that were at times erroneous. The newly generated Kenyan sequences described above were then added to this dataset.

To remove sequence duplicates, the sequences were binned by country of sampling, filtered of duplicates and then re-collated into a single dataset. Alignments were generated using MAFFT and edited manually in AliView and Geneious. For each RSV group, four sequence datasets with varying sequence lengths were prepared (*Table 5.1* and *Figure 5.1*); (i) complete genomes [>14,000 bases], (ii) complete G gene [>900 bases], (iii) Partial G (1st hypervariable regions through to terminal stop codon; 650-700 bases), and (iv) Partial G (2nd hypervariable region only; 300-350 bases).

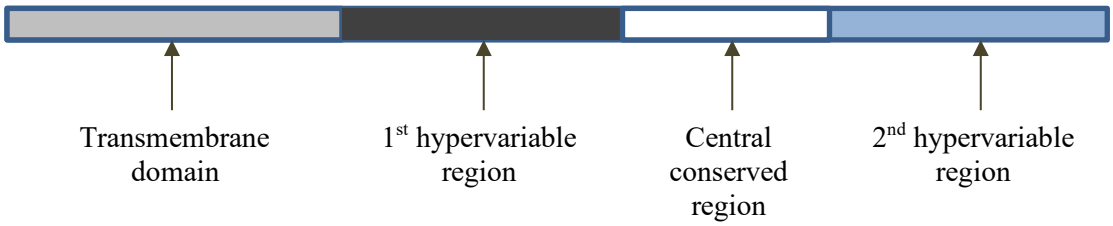


Figure 5.1: Schematic of the functional domains of the G protein

Table 5.1: Datasets available for local and global phylogeographic analysis

	RSV-A				RSV-B			
	Full genomes	Full G (>900nt)	Partial G (>650-700nt)	Partial G (300-350nt)	Full genomes	Full G (>900nt)	Partial G (>650-700nt)	Partial G (300-350nt)
Local			1,346				1,180	
Global	728	1,402	1,892	3,632	466	836	1,528	3,626

Centroid locations (geographical centre of the country in latitude-longitude co-ordinates) were downloaded to generate an inter-location great-circle distance matrix (https://worldmap.harvard.edu/data/geonode:country_centroids_az8). We obtained air transportation and model-based fluxes from the global epidemic and mobility (GLEAM) team (Broeck *et al.* 2011). The GLEAM model integrates real demographic and mobility data in a fully stochastic metapopulation network model and allows for the detailed simulation of the spread of infectious agents around the globe. In these simulations, the world population is divided into geographic census areas (subpopulations) that are defined around transportation hubs and connected by mobility fluxes (Balcan *et al.* 2009). Further, within each subpopulation the disease spreads between individuals while individuals can also move from one subpopulation to another along the mobility network according to high quality transportation data and thus simulating the global spreading pattern of epidemic outbreaks (Pastor-

Satorras *et al.* 2015). To model seasonal dynamics, the GLEAM framework was extended by using a compartmental model with the following parameters that best fit influenza seasonal dynamics (Poletto, private communication); (i) a temporary immunity to the virus with an average duration of between 2-10 years, (ii) a geographically dependent seasonal transmission in temperate areas, varying between a minimum basic reproductive number during summer ($R_{min} \in$ 0.5-0.75) and a maximum basic reproductive number during winter ($R_{max} \in$ 1.25-2.0), and (iii) a constant transmission with a basic reproductive number equal to R_{max} in the tropics (Poletto, private communication). In the compartmental model, individuals are divided in susceptible, latent, symptomatic infectious (that may or may not be travel dependent on the severity of symptoms), asymptomatic infectious and recovered (immune to the virus), with the average duration of the exposed and infection period set to 1.1 and 2.5 days, respectively. These GLEAM simulations were used as the 'flux' matrix predictors in the GLM-based phylogeographic analyses below.

Temporal signal

In order to visually examine the degree of temporal signal or accumulation of divergence in the datasets over the sampling time interval, the exploratory linear regression approach implemented in TempEst v1.5.1 (Rambaut *et al.* 2016) was employed. For each dataset, a maximum likelihood (ML) tree was generated using FastTree (Price, Dehal and Arkin 2010) under a generalized time-reversible (GTR) substitution model, and then plotted the root-to-tip divergences as a function of sampling time according to a rooting that maximizes the Pearson product moment correlation coefficient in TempEst. Outlier sequences that were either not divergent enough or too divergent, *Figure 5.2*, were removed from the datasets. The new

datasets were used to generate new ML trees that were heuristically time-transformed using TempEst and subsequently used as starting trees in phylogeographic analyses below to reduce burn-in of the Marko Chain Monte Carlo (MCMC) analyses.

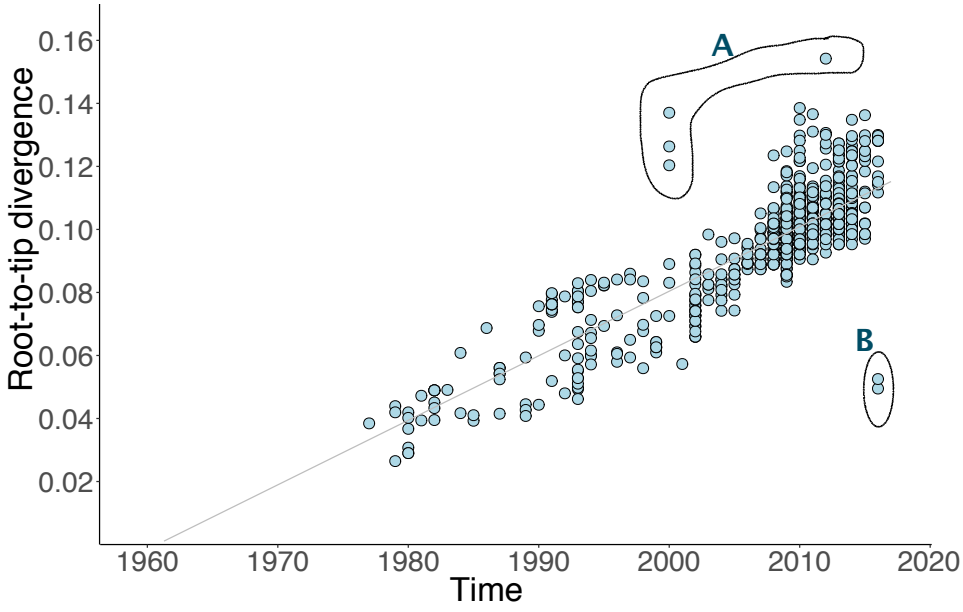


Figure 5.2: Examination of temporal signal in global sequence datasets in TempEst Root-to-tip divergence as a function of sampling time for ML tree clusters of global RSV-A full G gene sequence dataset highlighting (A) Too divergent and (B) Non-divergent sequences.

Discrete phylogeographic analysis: Kenya

The RSV-A and RSV-B partial G gene datasets used to perform this analysis included all the Kenyan sequences available in GenBank in addition to the new sequences generated in this study. Spatial diffusion dynamics among a set of six geographic locations in Kenya (i.e. Siaya, Kisumu, Kakuma, Nairobi, Dadaab and Kilifi; see Chapter Two *Figure 2.3*) was estimated using a Bayesian discrete phylogeographic approach (Lemey *et al.* 2009). Conditioning on the geographic locations of the tips of the G gene phylogenies, the transition history among the locations is modelled as a continuous time Markov chain (CTMC) process and thereby making inferences of the

unobserved locations at the ancestral nodes in each tree of the posterior distribution. A non-reversible CTMC model (Edwards *et al.* 2011) was used and incorporated Bayesian stochastic search variable selection (BSSVS) to identify a sparse set of transition rates that adequately summarize the epidemiological connectivity between the locations (Lemey *et al.* 2009). All the Markov chain Monte Carlo (MCMC) sampling analyses were performed using BEAST in conjunction with BEAGLE library to enhance computation speed (Ayres *et al.* 2012; Drummond *et al.* 2012). Two independent MCMC chains of 100 and 200 million steps were ran, and then the log and tree files combined. TreeAnnotator (Drummond *et al.* 2012) was used to summarize the location estimates on MCC trees after discarding 10% of the trees as chain burn-in. The MCC trees with annotations were then visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Building up on previous work from Kilifi (Agoti *et al.* 2014b; Otieno *et al.* 2017), a separate analysis on the Kenyan RSV genotype ON1 viral evolution and dispersal was conducted using BEAST as described in Methods section in Chapter Four.

Discrete phylogeographic analyses: Global

We used a GLM extension of the discrete phylogeographic model in BEAST (Faria *et al.* 2013; Lemey *et al.* 2014), which not only aims to estimate the relative intensities of viral dispersal among pairs of locations but also determine the subset of explanatory variables that help to explain such intensities. The viral diffusion rates are modelled as a log linear function of the selected predictor variables. The support and effect size for each predictor is estimated using inclusion probabilities and GLM coefficients, respectively (Lemey *et al.* 2014). The predictors of spread for the GLM

analysis included the great-circle distance between location pairs, air traffic data and GLEAM seasonal flux simulations. All the predictors were log transformed prior to their inclusion in the GLM analyses.

Prior to setting up the GLM analyses, to determine how best to include the predictors for the three traits, two test analyses were performed on the RSV-B G gene dataset; one with all the three predictors for countries and two predictors (air travel and fluxes) for the other two subdivisions, and one with two predictors (air travel and fluxes) for all the three subdivisions. Since distance was not preferred at the more fine (country) spatial subdivision level in addition to the difficulty in defining a distance between larger areas (where a centre becomes a very crude approximation), it was not included as a predictor in the continental and hemisphere traits. The XML files were edited to change the standard estimation procedure whereby instead of allowing all predictors to be included and excluded from the model, only one predictor was included. For this heterogeneous and fragmentary sampling, we were not concerned with how much better the predictors were than the 'equal-rates' null model but wanted to test which predictors suited best to protect against sampling bias. In addition, we did not want air travel and fluxes to both get into the model as fluxes are derived from air travel. Prior inclusion probabilities were specified that put 50% or 33% prior probability on no predictor being included, for the two and three predictor analyses respectively, and a normal prior with a mean of 0 and a standard deviation of 2 on the coefficients in log space. Bayes factor (BF) support for predictors was calculated based on the ratio of posterior to prior odds for predictor inclusion.

These analyses were initially performed with two of the four global datasets with the best molecular evolutionary signals, i.e. the complete G gene and whole genomes. For each dataset, in addition to the country of sampling, two more coarse spatial subdivisions were assigned, i.e. continent and hemisphere. There were six states for the continent trait (North America, South America, Africa, Asia, Europe, and Australia and Oceania) and three states for the hemisphere trait (temperate Northern, Tropics [including the subtropical regions] and temperate Southern; https://en.wikipedia.org/wiki/File:World_map_indicating_tropics_and_subtropics.png. We also estimated the number of location transitions (Markov jumps) and the time spent in each location state (Markov rewards) in order to have a complete summary of the spatial dispersal processes (Minin and Suchard 2008).

5.4 Results

5.4.1 ***G-gene sequencing and assembly***

Of the 400 CDC-K countrywide samples (2011-2014), 372 (93%) were successfully amplified out of which 327 (87.9%) were successfully sequenced. Since only the non-group specific outer primers were used for both G-gene amplification and sequencing, both group A and B sequences from six of these samples were assembled and thereby assembling a total of 333 sequences, *Table 5.2*. These G gene sequences were 518-824nt in length and covered a portion of the first hypervariable region through to the end of the G protein.

Table 5.2: Number of sequenced Kenyan non-Kilifi RSV-A and RSV-B samples (2011-2012) by group, year and location

Site	RSV-A				RSV-B			
	2011	2012	2013	2014	2011	2012	2013	2014
Dadaab	1	0	0	1 [1]$^{\alpha}$	60	0	0	3
Siaya	32	25 [4]	20 [8]	0	56	13	5	0
Kakuma	31	8	9 [5]	1	52	6	8	2
Total	**64**	**33**	**29**	**2**	**168**	**19**	**13**	**5**
	128				**205**			

$^{\alpha}$ Number of ON1 sequences

The sampling over the years was quite sparse for Dadaab in this dataset, except for 2011 where there were 60 group B sequences. However, with regard to sequence availability in GenBank, these are the first set of RSV sequences from Siaya and Kakuma. Therefore, they make an important contribution to available sequences than can be used to understand the local epidemiology and molecular evolution of the virus. In addition to Kilifi, a total of 18 ON1 viruses from the rest of Kenya were sequenced; 12 from Siaya, 5 from Kakuma and 1 from Dadaab.

5.4.2 ***Sampling bias in local and global datasets***

There was tremendous heterogeneity in sampling distribution between countries, *Appendices 7.12* and *7.13*, for both the G and the whole genome datasets. Even the local sampling of viruses in Kenya was heterogenous with majority of the sequences in the group A (77.7%) and B (73.1%) datasets collected from Kilifi, *Table 5.3*. Measuring the heterogeneity with Shannon entropy, however, yielded no obvious disparities (*Appendices 7.12* and *7.13*). There were no obvious disparities either in the Shannon entropy measure of heterogeneity with the larger geographic subdivisions of

continents and hemispheres, even though Australia and Oceania and the Southern hemisphere consistently had much lower numbers in the respective datasets. Nonetheless, analysis of the datasets was undertaken while expecting some challenges in getting an accurate and detailed information on the local and global RSV dissemination with the heterogenous sampling.

Table 5.3: Number of RSV sequences available for phylogeographic analysis from across Kenya, 1999-2016

Site	RSV-A	RSV-B
Dadaab	173 [2008-2011,2014]$^{\alpha}$	169 [2007-2011]
Kilifi	1046 [2000-2015]	863 [1999-2016]
Nairobi	6 [2011]	8 [2011]
Kisumu	3 [2011]	-
Kakuma	46 [2011-2014]	67 [2011-2014]
Siaya	72 [2011-2013]	73 [2011-2013]
Total	**1,346**	**1,180**

$^{\alpha}$The dates (years) of samples collection

5.4.3 ***Estimating the date of introduction and evolutionary rate of Kenyan RSV genotype ON1 viruses***

The Kenyan ON1 sequence dataset used in this analysis comprised 378 sequences, of which 360 were from Kilifi, 12 from Siaya, 5 from Kakuma, and 1 Dadaab. While the first ON1 virus isolated in Kilifi was collected in February 2012 (Agoti *et al.* 2014b), from this dataset the earliest non-Kilifi ON1 viruses were collected in April 2012 from Siaya, July 2013 in Kakuma and May 2014 in Dadaab (*Figure 5.3*). The most recent common ancestor (MRCA) for the ON1 viruses in Kenya was dated December 2010 [95% Highest Posterior Density (HPD): December 2009 – October 2011], about

one year and two months before initial detection in Kilifi. These ON1 viruses had an estimated substitution rate of 4.21×10^{-3} [95% HPD: 3.13×10^{-3} – 5.51×10^{-3}] substitutions per site per year which was quite similar to a previous estimate using a global ON1 G-gene dataset [4.10×10^{-3}] (Duvvuri *et al.* 2015).

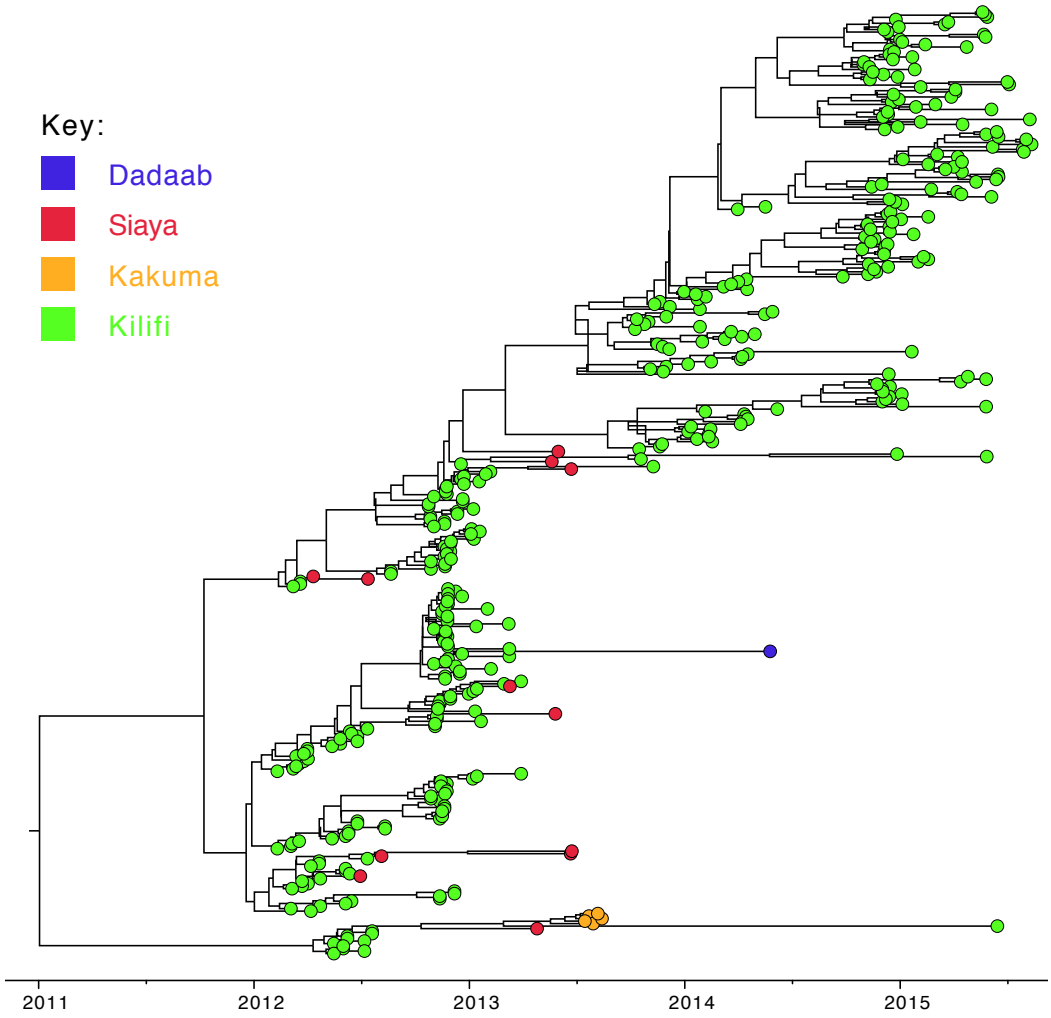


Figure 5.3: Time-calibrated MCC tree inferred for 378 G gene RSV genotype ON1 sequences from Kenya

The tips of the tree are coloured according to the location of sample collection as shown by the key at the top left.

5.4.4 ***Local dispersal of RSV in Kenya***

To capture the underlying RSV spatial diffusion dynamics across Kenya, the best supported rates of discrete location transitioning among all pairs of locations as inferred using BSSVS were summarized in *Figure 5.4* and *Appendix 7.14*. For RSV-A and RSV-B, sequences were available from six and five locations, respectively. The best supported virus transition rates originated from Kilifi and destined for both geographically close locations such as Dadaab in Eastern Kenya and far flung areas such as Kakuma in the North West, an observation that most likely reflects the sampling bias in the datasets as shown in *Table 5.3*. However, locations that are in relatively close proximity from each other (e.g. Siaya and Kakuma or Kilifi and Dadaab) had well-supported diffusion pathways for both RSV groups. The location state estimates obtained by discrete phylogeographic reconstruction revealed a weakly spatially structured Kenyan RSV population (*Figure 5.5*).

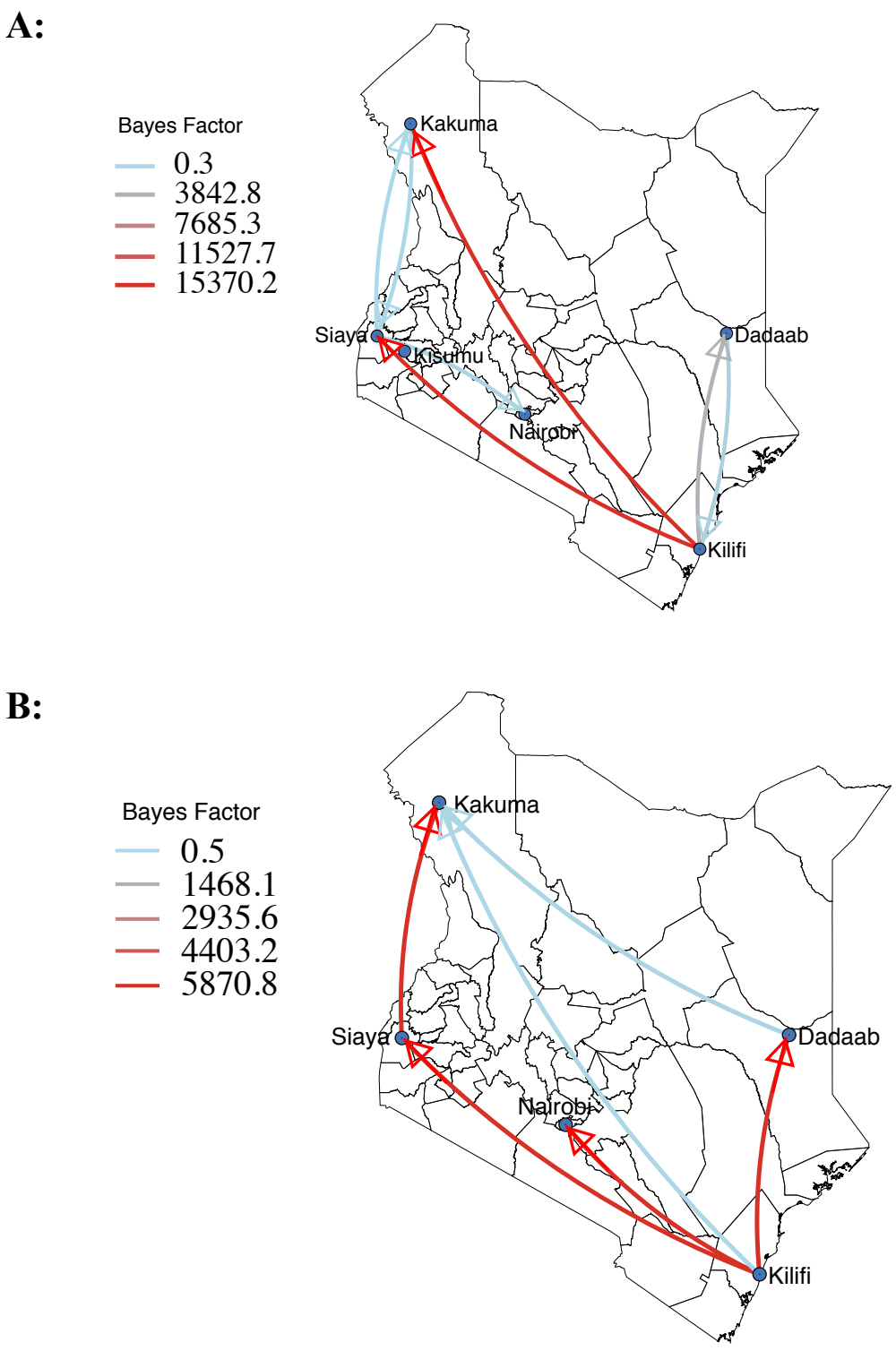
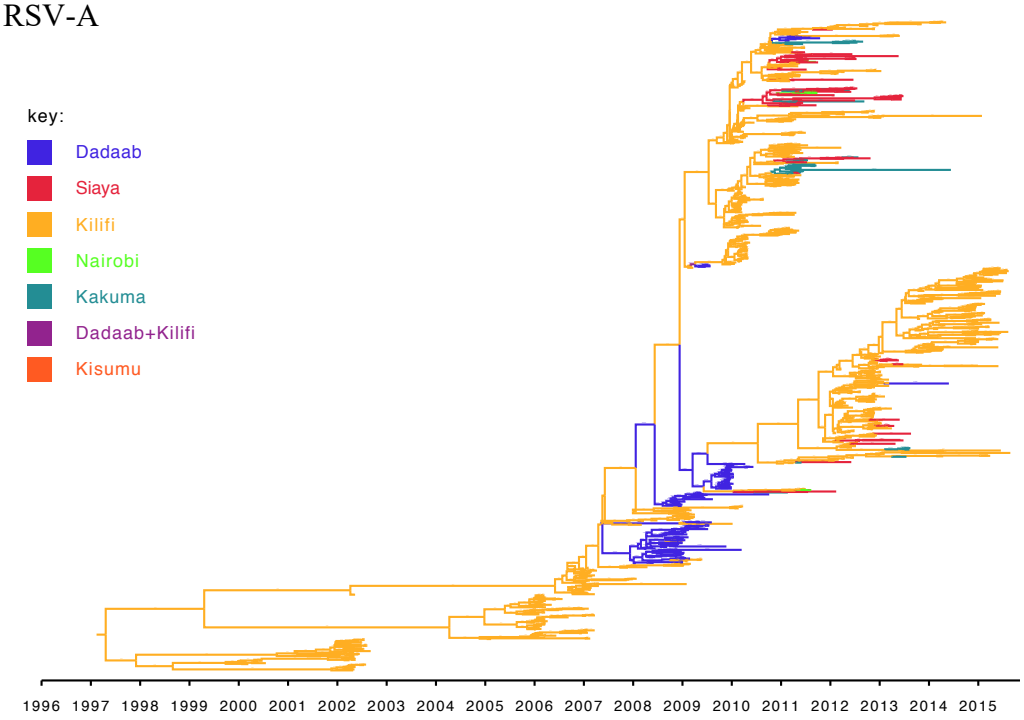


Figure 5.4: Bayes Factor (BF) support for RSV spatial diffusion in Kenya

Panel (A) RSV-A and (B) RSV-B. Rates are shown for BF with posterior probability >0.9. The line color represents the relative strength by which the rates are supported: light blue and red reflect relatively weak and strong support, respectively.

A: RSV-A



B: RSV-B

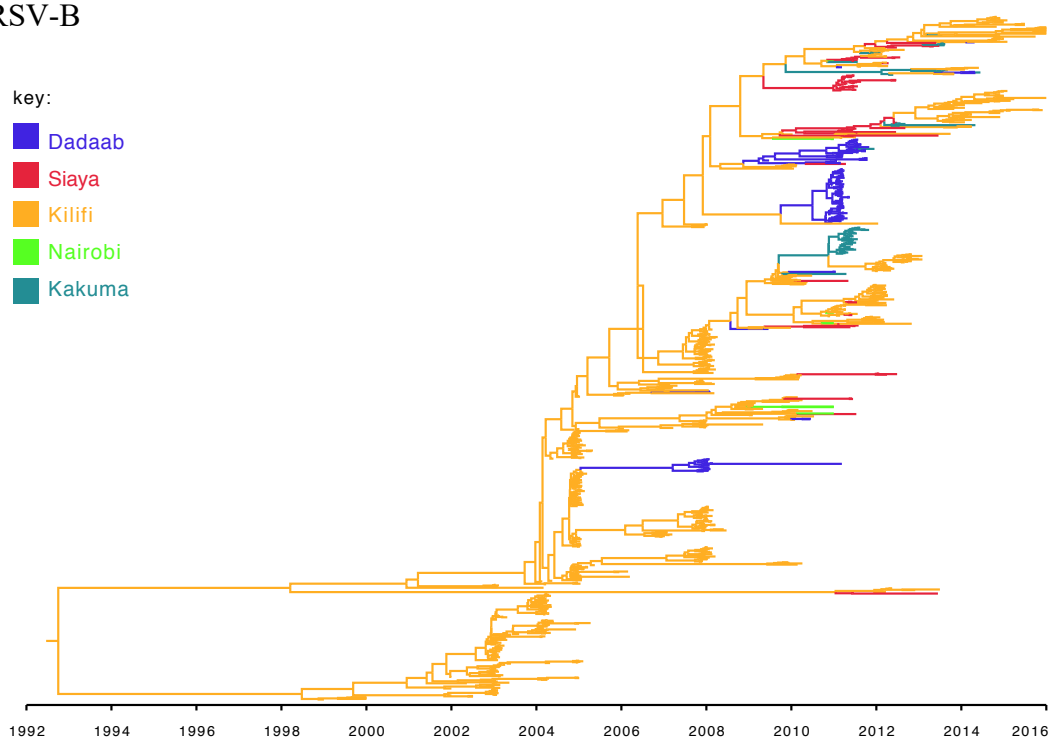


Figure 5.5: Time-calibrated MCC trees inferred for Kenyan partial G gene sequences of RSV-A and RSV-B

Branches are coloured according to the most probable location state, indicated in the coloured key at the top left.

5.4.5 ***Global dispersal of RSV***

In this analysis, two sequence datasets (full G gene and whole genome) per RSV group were used to infer the spatial dispersal of RSV between discrete locations/regions by summarizing the estimated virus jumps between countries, continents and hemispheres. In general, there were relatively higher jumps between countries using the WGS datasets than the G gene datasets, with the jumps mostly transcontinental but with fewer countries represented compared to the G gene (*Figure 5.6*). With the G gene datasets, *Figure 5.7*, the best export destinations per country was often to a nearby country and within the same continent while most imports into countries mostly arose from the US, China, and India that have large land masses and huge populations.

At the continental level, *Figures 5.8* and *5.9*, RSV-A G, RSV-A WGS and RSV-B G all supported Asia and North America as the most likely sources of RSV strains. These results mirrored those from the country trait above where the US (North America), India (Asia) and China (Asia) were the predominant exporters of viruses into other countries. In contrast, RSV-B WGS indicated that Europe was the predominant exporter of RSV-B viruses.

Finally, at the hemisphere level, the highest jumps were between the Northern hemisphere and the Tropics with the two regions seemingly seeding each other (*Figure 5.10*). Virus importation into the Southern hemisphere arose from the

Northern hemisphere and the Tropics but with minimal exportations into these source populations. These observations were consistent across all the four datasets.

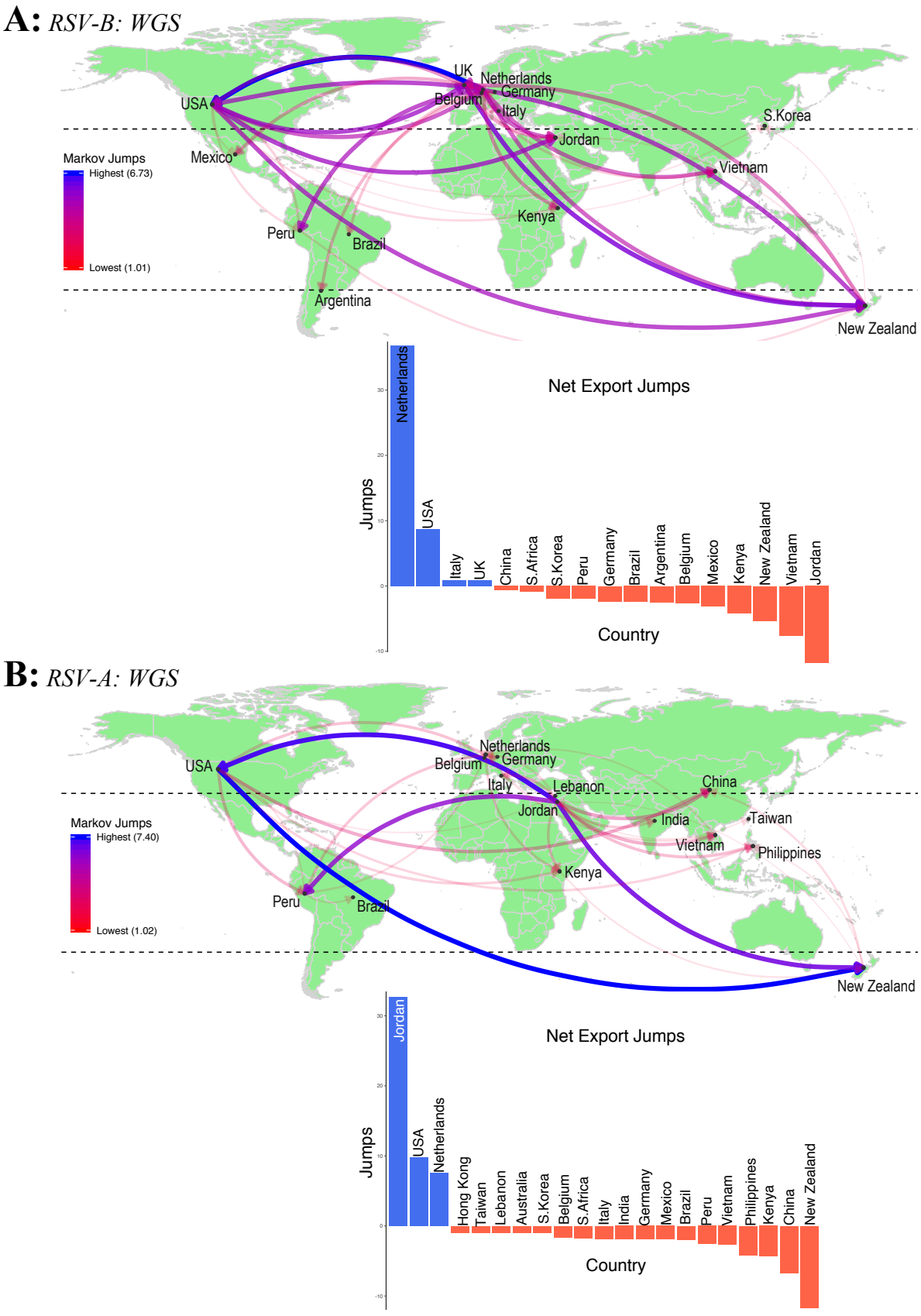


Figure 5.6: WGS based maps and bar graphs of the global RSV-A and RSV-B pairwise Markov jump history for the country trait in the posterior tree distribution.

The thickness and colour of the connecting lines are based on the number of jumps.

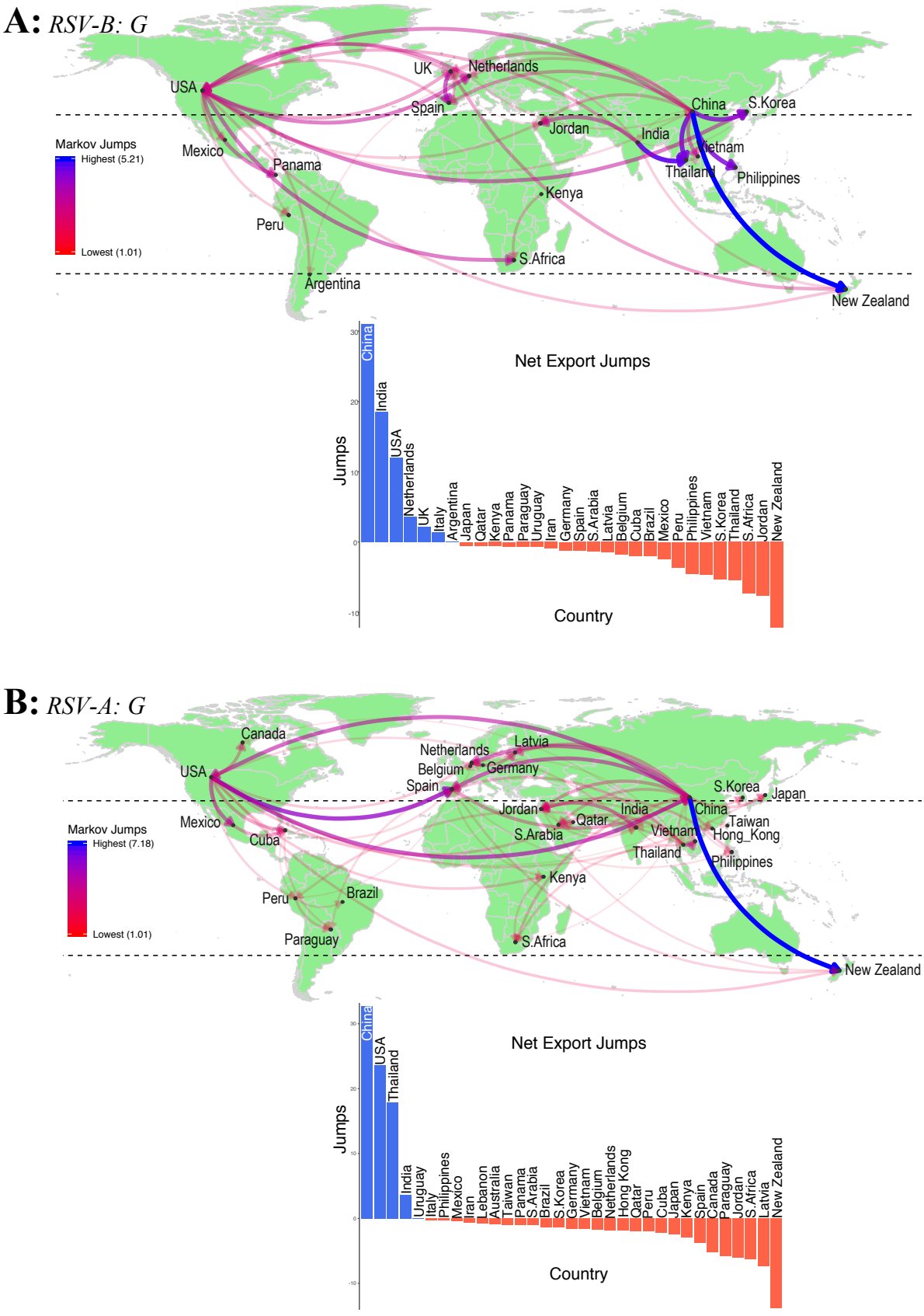


Figure 5.7: G gene based maps and bar graphs of the global RSV-A and RSV-B pairwise Markov jump history for the country trait in the posterior tree distribution.

The thickness and colour of the connecting lines are based on the number of jumps.

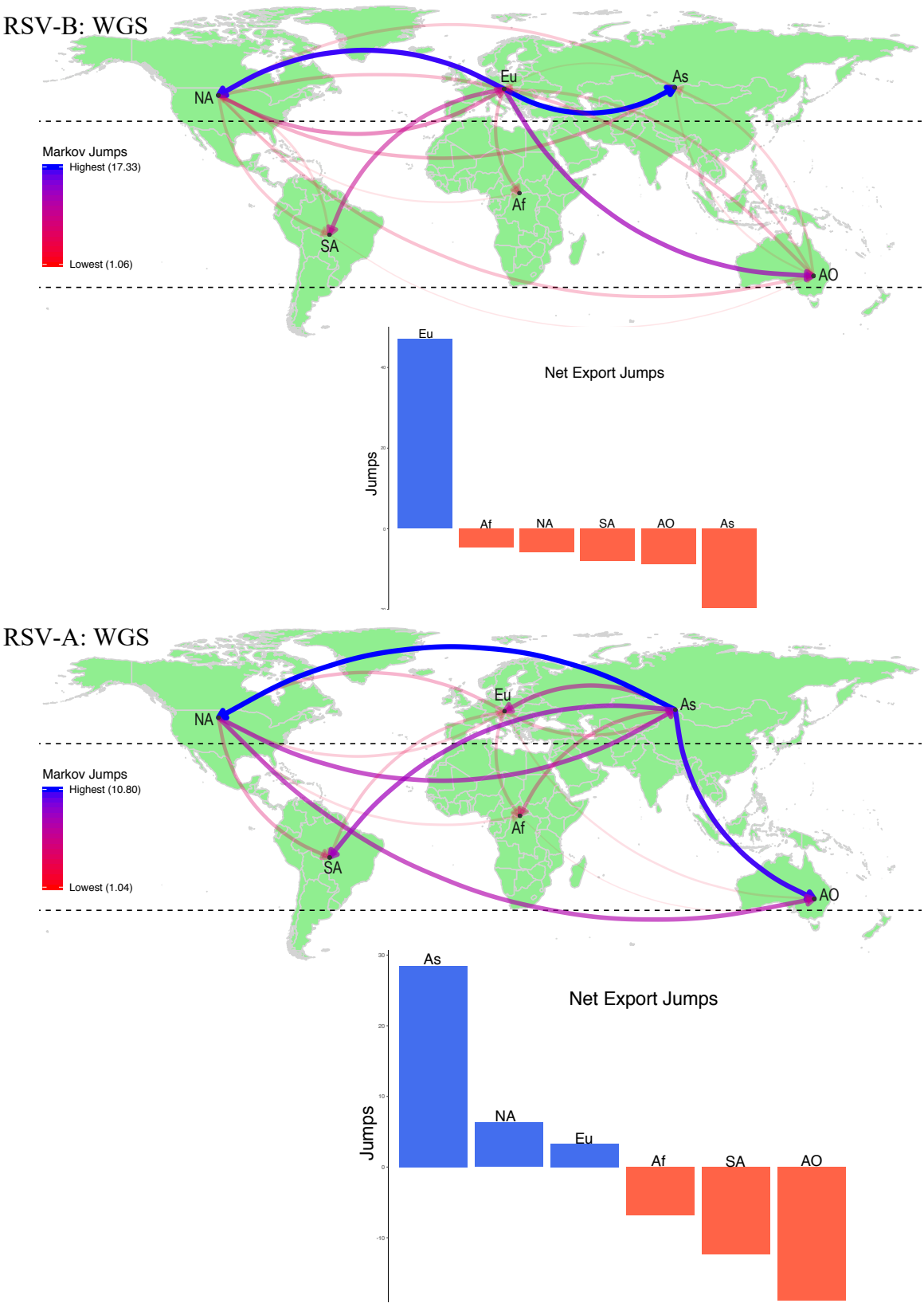


Figure 5.8: WGS based maps and bar graphs of the global RSV-A and RSV-B pairwise Markov jump history for the continent trait in the posterior tree distribution.

The thickness and colour of the connecting lines are based on the number of jumps. As: Asia, NA: North America, SA: South America, Eu: Europe, AF: Africa, AO: Australia and Oceania

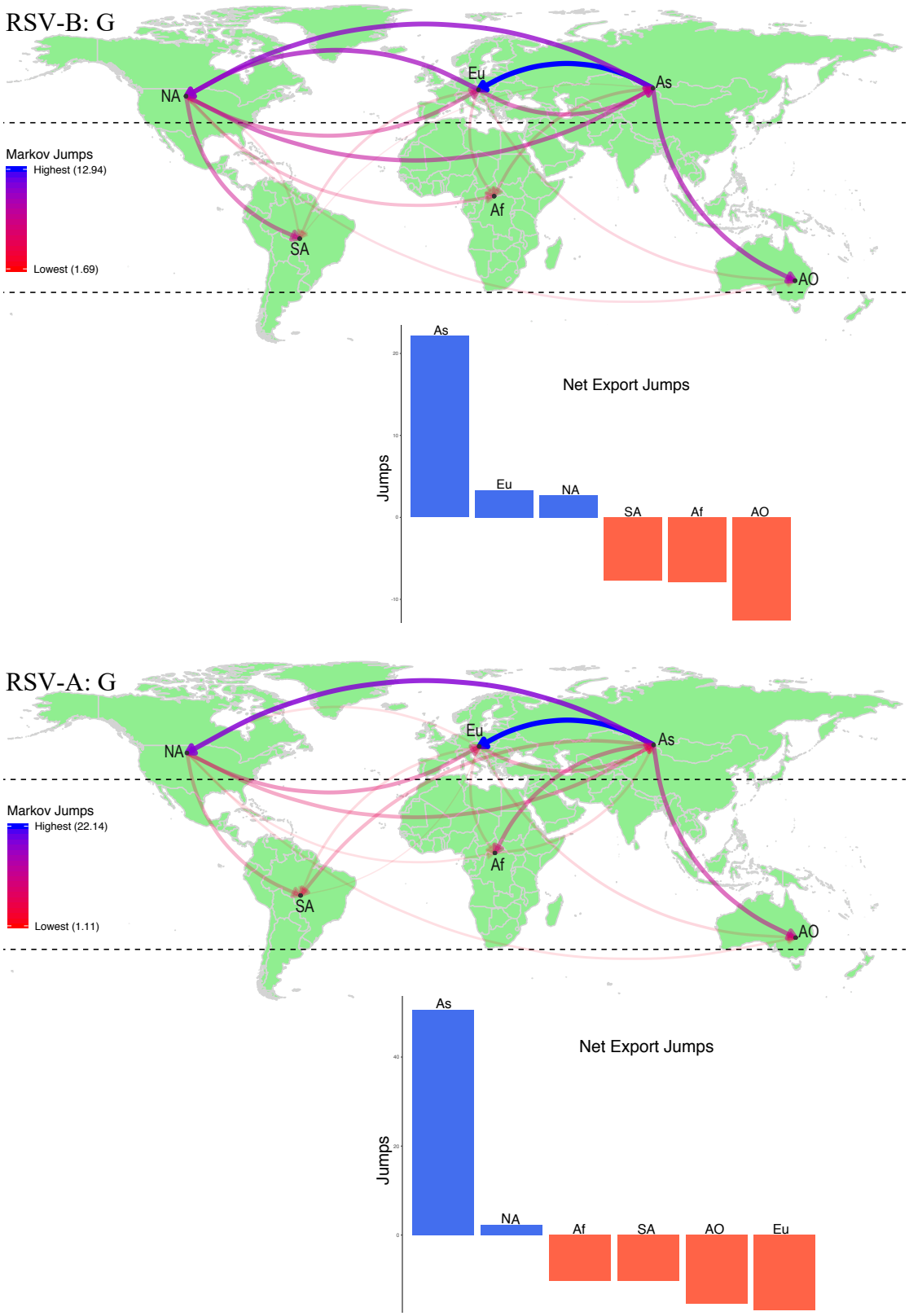


Figure 5.9: G gene based maps and bar graphs of the global RSV-A and RSV-B pairwise Markov jump history for the continent trait in the posterior tree distribution.

The thickness and colour of the connecting lines are based on the number of jumps. As: Asia, NA: North America, SA: South America, Eu: Europe, AF: Africa, AO: Australia and Oceania

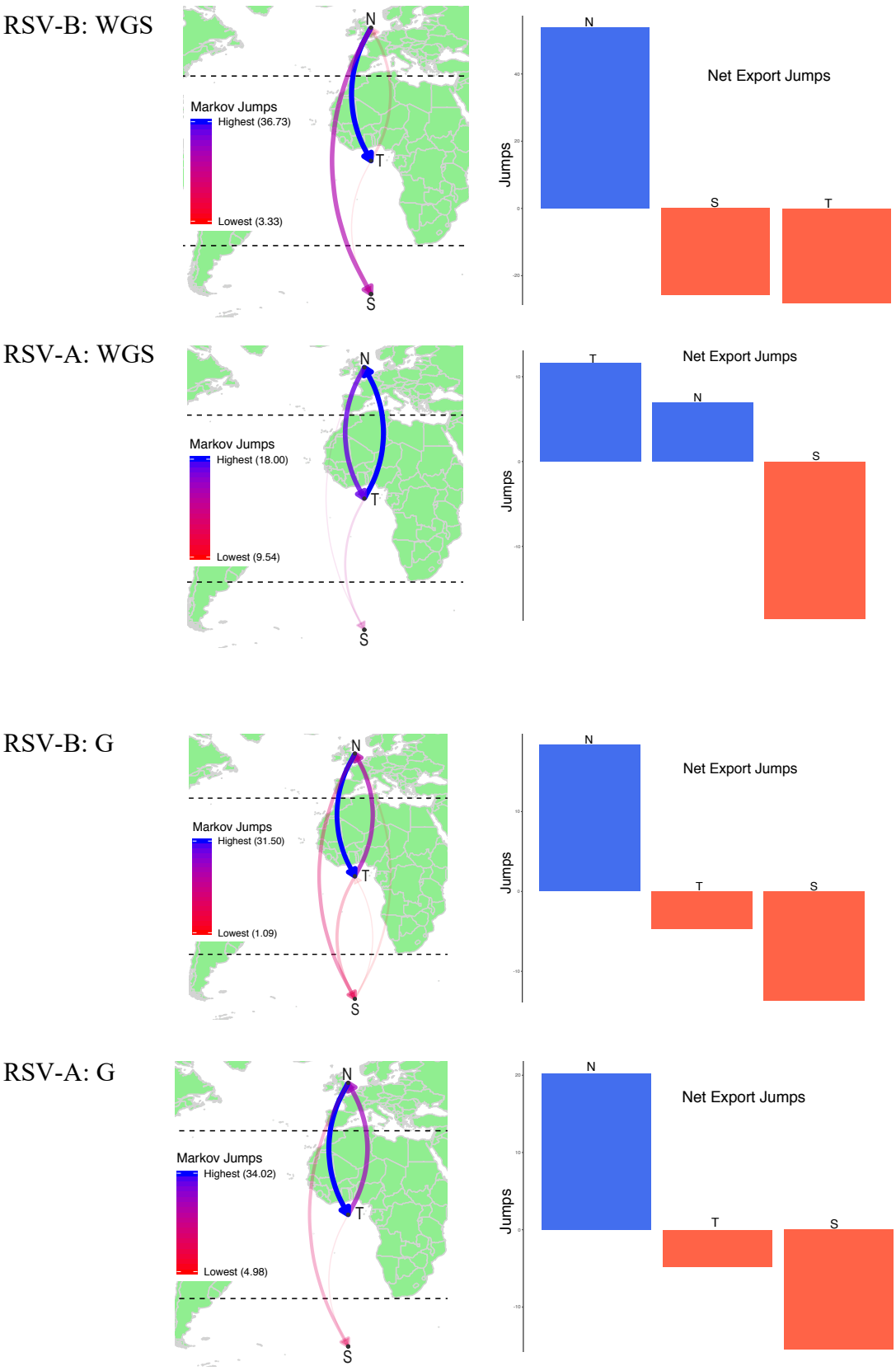


Figure 5.10: Maps and bar graphs of the global RSV-A and RSV-B pairwise Markov jump history for the hemisphere trait in the posterior tree distribution.

The thickness and colour of the connecting lines are based on the number of jumps. N: Northern hemisphere, T: Tropics, S: Southern hemisphere.

5.4.6 ***Predictors of global RSV dispersal***

The analyzed predictors of global RSV spread and the estimated inclusion probabilities and Bayes factors (BF) support for discrete diffusion rates are shown in *Table 5.4*. For RSV-B, the best predictor of RSV spread between countries was the seasonal fluxes while air traffic was the best predictor of spread between the larger spatial sub-divisions of continents and hemispheres. These observations were consistent for both the G gene and whole genome datasets. For RSV-A, the seasonal fluxes were the best predictors of RSV spread both between countries and continents. However, there was some inconsistency in the best predictor of RSV-A spread between the G gene and WGS datasets at the hemisphere level; for RSV-A G gene, air traffic was most preferred while for RSV-A WGS the seasonal fluxes were preferred. While the reason for this inconsistency between the RSV-A G gene and WGS datasets at this spatial subdivision was not immediately clear, the only noticeable difference was that there were more samples from the Tropics compared to those from the Northern hemisphere for the RSV-A WGS dataset unlike for the RSV-A G, RSV-B G, and RSV-B WGS datasets where there were more samples from the Northern hemisphere than the Tropics. Nonetheless, often there wasn't a great distinction between the predictors with air traffic and the fluxes receiving similar support and this may in part be due to the fluxes being derived from air-travel.

Table 5.4: Predictors of global RSV spread between countries, continents and hemispheres

RSV-B				
	Country			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	0.2396	0.6399	0.4168	1.4508
Distance	0	0	0.1160	0.2664
Flux	**0.7604**‡	**6.4418**	**0.4673**	**1.7807**
	Continent			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	**0.5844**	**1.4064**	**0.8141**	**4.3791**
Flux	0.4156	0.7111	0.1859	0.2284
	Hemisphere			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	**0.5971**	**1.4818**	**0.7175**	**2.5403**
Flux	0.4029	0.6749	0.2825	0.3937

RSV-A				
	Country			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	0.3375	1.0341	0.1226	0.2838
Distance	0	0	0.1016	0.2295
Flux	**0.6625**	**3.9862**	**0.7758**	**7.0257**
	Continent			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	0.2899	0.4082	0.4007	0.6685
Flux	**0.7101**	**2.4499**	**0.5993**	**1.4958**
	Hemisphere			
	G		***WGS***	
Predictor	*Inclusion Probability*	*Bayes Factor*	*Inclusion Probability*	*Bayes Factor*
Air	**0.5989**	**1.4931**	0.4545	0.8332
Flux	0.4011	0.6698	**0.5455**	**1.2002**

‡ In bold, the best predictor of spread for each dataset and spatial sub-division

5.5 Discussion and Conclusions

RSV is the most important cause of viral ALRI in children worldwide and a health concern in terms of morbidity, mortality and costs (Simoes 1999; Cane 2001; Shi *et al.* 2017). Persistence of the virus in communities, characterized by annual, biannual and biennial epidemics, is thought to be driven in part by frequent introductions of virus variants from other communities (Weber, Mulholland and Greenwood 1998; Centers for Disease Control and Prevention (CDC) 2004; Mlinaric-Galinovic *et al.* 2012; Agoti *et al.* 2014a, 2015b; Otieno *et al.* 2016). In this study, we present novel work on understanding the local and global spread of RSV by reconstructing the phylogeographic history of the virus in a discrete (Lemey *et al.* 2009) space using Bayesian inference, and test and quantify a range of potential predictors of spatial spread (Lemey *et al.* 2014).

We found that the local RSV dispersal in Kenya is predominantly characterized by virus transitions from Kilifi into other locations within the country, even though this observation is biased by heavy sampling of viruses from Kilifi. However, locations in close proximity to each other also had well-supported diffusion rates for both RSV groups implying regional circulation of viruses may be important for virus persistence within the country. The notion of sub-national transmission may be supported by the observation of weak spatial structuring of the virus populations. It is known from previous studies in Kilifi that virus persistence in this community is modulated by frequent introduction of virus variants from outside the community (Agoti *et al.* 2015b; Otieno *et al.* 2016). The unanswered question that would greatly benefit understanding the epidemiology of RSV in the country is whether the introduction of a virus variant(s), e.g. ON1, is characterized by (i) entry into a single location (e.g.

Kilifi) which then spread across the country and subsequently maintained through sub-national transmission, or (ii) entry into multiple locations/regions independently with some exchange of variants between regions (*Figure 1.5*).

In the global context, the northern hemisphere was predicted to be the major source population of RSV into the tropics and the southern hemisphere. This is at odds with the influenza virus where the tropics (mostly East and Southeast Asia, and India) has been shown to be the source population from incidence data and phylogeographic analyses (Russell *et al.* 2008; Lemey *et al.* 2014; Bedford *et al.* 2015). If one compares the air traffic and fluxes summarized by hemisphere, the main difference is the asymmetry in the fluxes, specifically between the tropics and the other two hemispheres. As the model underlying the fluxes allow continual transmission in the tropics similar to influenza, this region acts as a source population in the fluxes. So, perhaps, (i) a source population in the tropics is not really the case for RSV or (ii) sampling bias prevents us from identifying the source. There were more samples from the northern hemisphere compared to the tropics, which would make it difficult to identify the tropics as a source. In the case of the RSV-A WGS dataset where there were almost twice as many samples from the tropics than the northern hemisphere, the number of virus jumps from the tropics to the northern hemisphere only marginally surpassing those from the northern hemisphere to the tropics. If the tropics were the source, then the difference in jumps might have been substantial. However, sampling bias isn't just about the difference in the number of samples as the distribution of the samples across time is equally critical and needs careful consideration in such analyses.

It has been reported that annual RSV outbreaks generally occur in the late fall and winter in the temperate regions and during the rainy seasons in the tropical regions (Moura *et al.* 2006; Goddard *et al.* 2007; Meerhoff *et al.* 2009; Murray *et al.* 2012; Obando-Pacheco *et al.* 2018). The study by *Obando-pacheco et al*. (Obando-Pacheco *et al.* 2018) reported that the RSV wave started between March and June in most countries in the Southern hemisphere and between September and December in the Northern hemisphere. There was a decrease in RSV activity from August to October in the Southern hemisphere and from February to May in the Northern hemisphere. However, they only partitioned the global countries into the Northern and Southern hemispheres unlike in this analysis where there was the tropical regions. Nonetheless, the period from mid-August to early October had the least RSV activity globally, a time that preceded the start of RSV activity in the Northern hemisphere. One can perhaps infer that annual RSV activity begins in the Northern hemisphere, and if this is the case then new variants initially circulate in the temperate north before onward transmission into the tropics and temperate south, a hypothesis that might agree with the observations from our phylogeographic analysis.

Phylogeographic analyses not only allow the description of the spatiotemporal patterns of viral spread but may also aid the formulation of hypotheses about the underlying processes that shape the dynamics of spread. It was observed that for RSV-A there is support for the seasonal fluxes as the best predictor of RSV spread between countries, but as soon as larger groupings such as continents and hemispheres are considered then air travel is favoured. This seems reasonable as the fluxes will strongly prefer transmission between countries within the same continent and hemisphere as they would have similar seasonal transmissibility (as modelled for

the fluxes). However, it is interesting that seasonal fluxes would still be preferred for the virus spread between continents and hemispheres for RSV-B. Whether this reflects an inherent difference in the epidemiology of the RSV-A and RSV-B viruses or simply arises from sampling bias warrants further investigation.

Sampling bias prevents us from making firm conclusions on the observations from this study. The sample patchiness in the datasets analyzed here arise from very heterogeneous RSV sequencing efforts across countries. These differences in the number of sequences do affect the migration rate estimates and hence the ancestral reconstructions. However, the current analysis takes an important first step towards the understanding of the dynamics and predictors of RSV spread at the local and global contexts. In the immediate future, the aim is to perform similar discrete spatial diffusion analyses using partial G gene sequence datasets (600-700nt) that capture more countries and samples, even though such a dataset runs the risk of reduced phylogenetic resolution due to shorter sequences. Due to the large numbers of these sequences and the computational time required for these analyses, the oversampled regions in these datasets will be down-sampled either (i) using known epidemiology (the prevalence in each location) or (ii) randomly for more similar or balanced numbers. Finally, there is a plan to quantify the virus spatial structure in the generated posterior set of trees by measuring the phylogenetic association in the location trait data (Wang *et al.* 2001; Parker, Rambaut and Pybus 2008; Lemey *et al.* 2009).

CHAPTER SIX

6 Overall Discussion

6.1 Introduction

Emerging infectious diseases (EIDs) are a significant burden on global economies and public health (Fauci 2001; Morens, Folkers and Fauci 2004). In fact, EID events have risen significantly over time (Jones *et al.* 2008). While public health surveillance of both old and new pathogens continues to rise, it still is inadequate. This is exemplified by recent reports showing unnoticed detection of pathogens such as Ebola and zika viruses for several months within populations (Dudas *et al.* 2017; Faria *et al.* 2017; Grubaugh *et al.* 2017; Metsky *et al.* 2017). Increased globalization and travel between states, countries and across continents has heightened and extended the potential of rapid pathogen spread and geographic reach, respectively, a risk that seems to be perennially underestimated (Gostin 2017). While RSV does not leave behind a sudden acute "death trail" as other viruses such as Ebola and SARS, new variants continue to emerge and the cumulative annual deaths and disease burden is of great significance especially in low and middle-income countries (Trento *et al.* 2003; Eshaghi *et al.* 2012; Shi *et al.* 2017). Therefore, increased surveillance and identification of new variants is important for the understanding the molecular epidemiology of RSV and design of future control strategies.

Pathogen genetic data contains information about spatio-temporal spread that can be extracted using phylodynamic approaches (Lemey *et al.* 2014; Faria *et al.* 2017). In addition, pathogen genome analyses can also shed light on the origins, evolution and transmission dynamics of a pathogen during an epidemic or across multiple epidemics (Smith *et al.* 2009; Holmes *et al.* 2016). Study patients' profiles can also be used to

investigate the clinical and demographic impact of emergent strains relative to previous or existing strains.

In this study, using the emergent RSV-A genotype ON1 as a unique tag, an integrated approach was undertaken using partial and whole genome sequence data, patient clinical and demographic profiles, human mobility (air traffic) data, and stochastic models of RSV disease transmission to understand; (1.) how new RSV variants are introduced and spread into communities, (2.) how they persist locally within those communities, (3.) the genomic signatures that differentiate emergent from existing variants and whether such substitutions may impact on fitness, and (4.) the patterns and drivers of RSV spread across geographically defined regions (local and global).

6.2 Key research findings

The key research findings from this thesis project are presented hereafter under each of the three major analyses conducted.

6.2.1 ***Molecular epidemiological, clinical and demographic characteristics of RSV-A genotype ON1 in Kilifi: analysis of G gene sequences***

In February 2012, the novel RSV group A genotype ON1 with the characteristic 72 nucleotide duplication within the G-gene was detected in Kilifi County, coastal Kenya. To better understand the molecular epidemiological characteristics of new RSV variants into a community, a genetic and phylogenetic analysis of a set of group A G-gene sequences from Kilifi (n=483) collected through the continuous surveillance of RSV-associated pediatric pneumonia admissions at KCH (2010/2011 to 2014/2015) was undertaken. The clinical and demographic information of the patients from whom these samples were collected were also statistically analyzed.

The following key findings were made:

1. Very rapid replacement of previously circulating genotype GA2 by ON1, quite unlike previous RSV-A genotype replacements within the same community. This might suggest that ON1 viruses are relatively fitter than the previous RSV-A genotypes. However, it was noted that the prevalence of ON1 varied globally and in some places the ON1 are seemingly not about to displace the predominating RSV-A genotype several epidemics after initial entry into such communities. We conclude that, in addition to genetic constitution, host and ecological differences may also determine the fitness advantage or success of an RSV variant.
2. Each RSV epidemic is characterized by the circulation of multiple variants, each differing sufficiently to suggest separate introductions into the community (as opposed to arising from diversification during the epidemic), very few of which persist across epidemics. While previous reports from this community had come up with the same conclusion analyzing RSV-A and B strains that had been circulating in Kilifi for a while and could be presumed to be the tendency of established variants (Agoti *et al.* 2015b; Otieno *et al.* 2016), this was a strong validation of these conclusions as the ON1 variant dynamics from initial introduction were analyzed. In all, it seems the concept of co-circulating lineages in communities exists at all levels of RSV classification (i.e. RSV group, genotype and genotype variants) albeit with varied rates of turnover year-on-year.
3. Accumulation of amino acid substitutions by the ON1 viruses, and more interestingly the acquisition of similar substitutions between adjacent and

corresponding positions within the duplicated region. This result probably gives the first indication of co-evolution between sites in RSV genotype ON1 viruses. While RSV uses CX3CR1 as a receptor to infect human ciliated airway epithelial cells, Hotard *et al* used cells expressing heparan sulphate to show that the duplication in the BA genotype resulted in better binding avidity (Hotard *et al.* 2015). Assuming the same effect in ON1 viruses and in ciliated airway epithelial cells, the longer attachment protein appears to offer more opportunities for variable changes hence greater diversity and potentially increased fitness over previous group A genotypes.

4. No clear evidence of altered pathogenicity of ON1 relative to GA2 in Kilifi, save for higher prevalence in inability to feed amongst ON1 infected children, with overall cases of very severe pneumonia equally prevalent in both genotypes. However, reports from around the globe give conflicting results with regard to ON1 RSV disease severity relative to other genotypes and could arise from methodological differences in analyses, clinical disease definitions and study designs, chance effects resultant from inadequate sample sizes, differences between viruses in different locations, or even host/environmental differences.

6.2.2 ***Whole genome evolutionary dynamics of RSV genotype ON1***

The G gene on its own has been shown, for example, to be occasionally insufficient in distinguishing between inpatient outbreak RSV strains isolated in a haematology-oncology and stem-cell transplant unit and outpatient epidemiologically-unrelated strains collected within the same time period (Zhu *et al.* 2017), and “who acquires

infection from whom" (Munywoki *et al.* 2014). A total of 184 RSV-A whole genome sequences (WGS) from Kilifi (Kenya) collected between 2011 and 2016, the first ON1 genomes from Africa and the largest collection globally from a single location, were generated and analyzed. WGS has the potential to provide a more detailed understanding of RSV molecular epidemiology, evolution, phylogeography, diagnostics and vaccine development.

What was novel and most interesting in this analysis was the identification of signature amino acid substitutions between Kilifi ON1 and GA2 viruses' surface proteins (G, F), polymerase (L) and matrix M2-1 proteins, some of which were positively selected, and thereby provided an enhanced picture of RSV-A diversity. Furthermore, five of the eleven RSV open reading frames (ORF) (G, F, L, N and P) formed distinct phylogenetic clusters for the two genotypes. This might suggest that coding regions outside of the most frequently studied G ORF also play a role in the adaptation of RSV to host populations or rather there could be linked selection (co-evolution) amongst antigenic and non-antigenic genes of novel RSV variants. However, without complementary and confirmatory functional analyses it is plausible that some of these signature substitutions could be neutral and provide no selective advantage.

Phylogenetic analysis re-affirmed the conclusions obtained by partial G-gene sequencing that RSV-A circulation in this coastal Kenya location is characterized by multiple introductions of viral lineages from diverse origins but with varied success in local transmission. Nonetheless, there have been questions on the nature of the emergence of ON1, i.e. if this occurred through a single ancestral ON1 virus or

multiple times. While it was noted that oddly placed viruses of a given genotype amongst viruses of another genotype might give the impression of multiple emergence, these viruses often had no descendants and came from the same labs, and therefore might be nothing more than cross-contamination and/or mis-assemblies.

This analysis reaffirms the epidemiological processes that define RSV spread, highlights the genetic substitutions that characterize emerging strains, and demonstrates the utility of large-scale WGS in molecular epidemiological studies.

6.2.3 ***Local and global RSV transmission dynamics***

In addition to understanding (i) how frequent introduction of new RSV variants contributes to its persistence within communities, (ii) the associated clinical and demographic impacts of such new variants, and (iii) the genomic signatures that characterize the new variants and may impact on fitness, it is equally important to understand the dynamics of the virus spread within and between geographically defined regions. Future public health control strategies might benefit from such insights on the dynamics of virus spread as they have the potential of improving the predictive models of disease control.

This phylogeographic analysis is a first for RSV in terms of aiming to unravel both the local and global patterns of RSV spread in discrete spaces. It was observed that virus spread between locations in close proximity might be important for virus persistence within the country. Favoured by more intensive sampling, the local RSV dispersal in Kenya was predominantly characterized by virus transitions from Kilifi into the other locations. Our hypothesis on the patterns of RSV spread in Kenya are

that a new virus variant(s) may be introduced into the country either at a single location then subsequently spread across the country or at multiple locations/regions independently with some exchange of variants between regions. With ongoing sequencing of more samples from across the country, it is expected that this might improve on the current observations and build a better picture of the local RSV spread.

For a well-studied virus like influenza, it has been accepted that the global source population of influenza diversity is the tropics and specifically mainland China and Southeast Asia. It is not known if (i) there is a specific source population of new RSV variants and (ii) where that population might be located. From this analysis, it is predicted that the northern hemisphere might be the major source population of RSV into the tropics and the southern hemisphere. In fact this observation might agree with the findings from *Obando-pacheco et al*. (Obando-Pacheco *et al.* 2018) where they gathered that the annual RSV epidemics globally began in the Northern hemisphere regions followed by the regions in the Southern hemisphere. However, the scope of the current analysis could not allow us to make inferences on why the Northern hemisphere could be the likely source population and is an area of interest for future analysis.

The phylogeographic analyses not only allowed us to describe the global spatiotemporal patterns of RSV spread but also formulate hypotheses about the underlying processes that shape the dynamics of spread. On one hand it was observed that seasonal fluxes and air travel were the best predictors of RSV-A spread at the fine (between countries) and coarse (between continents and hemispheres) geographic

scales, respectively. On the other hand, seasonal fluxes were the most preferred predictor of spread at all scales for RSV-B viruses. However, it is not apparent if this observation reflects on an inherent difference in the epidemiology of the RSV group A and B viruses or arises from sampling bias in our datasets.

Sampling heterogeneity in the analyzed datasets presents a challenge that prevents us from making firm conclusions about our observations. However, the current analysis makes the important first step towards the understanding of the dynamics and predictors of RSV spread both at the local and the global contexts. It might also spur discussions and concerted efforts, as is the case with influenza, on intensive surveillance and sequencing of RSV strains across the world.

6.3 Study limitations

While all the study objectives were addressed, sampling bias presented a great challenge in the analysis of the local and global patterns and drivers of RSV spread. For the local analysis, samples from Nairobi which we hypothesize to be crucial to understanding local spread patterns (based on geographic and administrative positioning) arrived in late November when this thesis was nearing submission and therefore sample processing, sequencing and analysis was not feasible. At the global level, there's an heterogenous effort in both surveillance and sequencing of RSV leaving such analyses fraught with inconclusive observations.

6.4 Thesis summary

Following initial detection of the genotype ON1 in Kilifi in 2012, there was rapid replacement of the previously circulating RSV group A genotype GA2 by ON1 in

subsequent epidemics. While this suggests elevated fitness of ON1 viruses, there was no clear evidence of altered pathogenicity of ON1 relative to GA2 in Kilifi. Signature amino acid substitutions were identified between surface proteins (G, F), polymerase (L) and matrix M2-1 proteins of Kilifi ON1 and GA2 viruses, suggesting co-evolution amongst antigenic and non-antigenic genes of RSV variants. Genetic and phylogenetic analyses reaffirmed previous conclusions that each RSV epidemic is characterized by the frequent introduction of multiple variants, few of which persist across epidemics. Finally, the phylogeographic analyses predicted the northern hemisphere to be the major source population of RSV into the tropics and the southern hemisphere and virus spread between locations in close proximity to be important for virus persistence within a country. Future work will explore (i) any functional consequences of the ON1-GA2 signature substitutions, and (ii) further RSV spread dynamics at the country and continental level with more sequence data currently being processed.

REFERENCES

Adazu K, Lindblade KA, Rosen DH *et al.* Health and demographic surveillance in rural western Kenya: A platform for evaluating interventions to reduce morbidity and mortality from infectious diseases. *Am J Trop Med Hyg* 2005;**73**:1151–8.

Agoti CN, Mayieka LM, Otieno JR *et al.* Examining strain diversity and phylogeography in relation to an unusual epidemic pattern of respiratory syncytial virus (RSV) in a long-term refugee camp in Kenya. *BMC Infect Dis* 2014a;**14**:178.

Agoti CN, Mbisa JL, Bett A *et al.* Intrapatient Variation of the Respiratory Syncytial Virus Attachment Protein Gene. *J Virol* 2010;**84**:10425–8.

Agoti CN, Munywoki PK, Phan MVT *et al.* Transmission patterns and evolution of respiratory syncytial virus in a community outbreak identified by genomic analysis. *Virus Evol* 2017;**3**, DOI: 10.1093/ve/vex006.

Agoti CN, Mwihuri AG, Sande CJ *et al.* Genetic Relatedness of Infecting and Reinfecting Respiratory Syncytial Virus Strains Identified in a Birth Cohort From Rural Kenya. *J Infect Dis* 2012;**206**:1532–41.

Agoti CN, Otieno JR, Gitahi CW *et al.* Rapid spread and diversification of respiratory syncytial virus genotype ON1, Kenya. *Emerg Infect Dis* 2014b;**20**:950–9.

Agoti CN, Otieno JR, Munywoki PK *et al.* Local Evolutionary Patterns of Human Respiratory Syncytial Virus Derived from Whole-Genome Sequencing. Perlman S (ed.). *J Virol* 2015a;**89**:3444–54.

Agoti CN, Otieno JR, Ngama M *et al.* Successive Respiratory Syncytial Virus Epidemics in Local Populations Arise from Multiple Variant Introductions, Providing Insights into Virus Persistence. *J Virol* 2015b;**89**:11630–42.

Ahmed A, Haider SH, Parveen S *et al.* Co-circulation of 72bp duplication group A

and 60bp duplication group B respiratory syncytial virus (RSV) strains in Riyadh, Saudi Arabia during 2014. *PLoS One* 2016, DOI: 10.1371/journal.pone.0166145.

Alonso WJ, Viboud C, Simonsen L *et al.* Seasonality of influenza in Brazil: A traveling wave from the amazon to the subtropics. *Am J Epidemiol* 2007;**165**:1434–42.

Anderson LJ, Bingham P, Hierholzer JC. Neutralization of respiratory syncytial virus by individual and mixtures of F and G protein monoclonal antibodies. *J Virol* 1988;**62**:4232–8.

Anderson LJ, Hendry RM, Pierik LT *et al.* Multicenter study of strains of respiratory syncytial virus. *J Infect Dis* 1991;**163**:687–92.

Anderson LJ, Hierholzer JC, Tsou C *et al.* Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies. *J Infect Dis* 1985;**151**:626–33.

Andrews S. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.

Auksornkitti V, Kamprasert N, Thongkomplew S *et al.* Molecular characterization of human respiratory syncytial virus, 2010-2011: Identification of genotype ON1 and a new subgroup B genotype in Thailand. *Arch Virol* 2014;**159**:499–507.

Avadhanula V, Chemaly RF, Shah DP *et al.* Infection with novel respiratory syncytial virus genotype Ontario (ON1) in adult hematopoietic cell transplant recipients, Texas, 2011-2013. *J Infect Dis* 2015;**211**:582–9.

Ayres DL, Darling A, Zwickl DJ *et al.* BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol* 2012;**61**:170–3.

Baele G, Lemey P, Bedford T *et al.* Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 2012;**29**:2157–67.

Baele G, Li WLS, Drummond AJ *et al.* Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Mol Biol Evol* 2013;**30**:239–43.

Baker M. De novo genome assembly: what every biologist should know. *Nat Methods* 2012, DOI: 10.1038/nmeth.1935.

Balcan D, Colizza V, Goncalves B *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc Natl Acad Sci* 2009;**106**:21484–9.

Balmaks R, Ribakova I, Gardovska D *et al.* Molecular epidemiology of human respiratory syncytial virus over three consecutive seasons in Latvia. *J Med Virol* 2014;**86**:1971–82.

Bankevich A, Nurk S, Antipov D *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 2012;**19**:455–77.

Becker Y. Respiratory syncytial virus (RSV) evades the human adaptive immune system by skewing the Th1/Th2 cytokine balance toward increased levels of Th2 cytokines and IgE, markers of allergy—a review. *Virus Genes* 2006;**33**:235–52.

Bedford T, Riley S, Barr IG *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* 2015, DOI: 10.1038/nature14460.

Belshaw R, Gardner A, Rambaut A *et al.* Pacing a small cage: mutation and RNA viruses. *Trends Ecol Evol* 2008;**23**:188–93.

Bigogo GM, Breiman RF, Feikin DR *et al.* Epidemiology of respiratory syncytial virus infection in rural and urban Kenya. *J Infect Dis* 2013;**208**:S207-16.

Blanc A, Delfraro A, Frabasile S *et al.* Genotypes of respiratory syncytial virus group

B identified in Uruguay. *Arch Virol* 2005;**150**:603–9.

Blanken MO, Rovers MM, Molenaar JM *et al.* Respiratory Syncytial Virus and Recurrent Wheeze in Healthy Preterm Infants. *N Engl J Med* 2013;**368**:1791–9.

Bliven KA, Maurelli AT. Evolution of Bacterial Pathogens Within the Human Host. *Virulence Mechanisms of Bacterial Pathogens, Fifth Edition*. American Society of Microbiology, 2016, 3–13.

Bloomquist EW, Lemey P, Suchard MA. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol Evol* 2010;**25**:626–32.

Bonfield JK, Smith K f, Staden R. A new DNA sequence assembly program. *Nucleic Acids Res* 1995;**23**:4992–9.

Borkje B. Diflunisal compared with naproxen in the treatment of osteoarthrosis of hip or knee. A double-blind trial. *Tidsskr den Nor Laegeforening* 1982;**102**:1774–93.

Bose ME, He J, Shrivastava S *et al.* Sequencing and analysis of globally obtained human respiratory syncytial virus a and B genomes. Varga SM (ed.). *PLoS One* 2015;**10**:e0120098.

Botosso VF, Zanotto PMD a, Ueda M *et al.* Positive Selection Results in Frequent Reversible Amino Acid Replacements in the G Protein Gene of Human Respiratory Syncytial Virus. Fouchier RAM (ed.). *PLoS Pathog* 2009;**5**:e1000254.

Brockmann D. Human Mobility and Spatial Disease Dynamics. *Reviews of Nonlinear Dynamics and Complexity*. Vol 2. 2010, 1–24.

Broeck W V.D., Gioannini C, Gonçalves B *et al.* The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infect Dis* 2011;**11**, DOI: 10.1186/1471-2334-11-37.

Brown IH, Harris PA, McCauley JW *et al.* Multiple genetic reassortment of avian and

human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype. *J Gen Virol* 1998;**79**:2947–55.

Bukreyev A, Yang L, Collins PL. The Secreted G Protein of Human Respiratory Syncytial Virus Antagonizes Antibody-Mediated Restriction of Replication Involving Macrophages and Complement. *J Virol* 2012;**86**:10880–4.

Bukreyev A, Yang L, Fricke J *et al.* The secreted form of respiratory syncytial virus G glycoprotein helps the virus evade antibody-mediated restriction of replication by acting as an antigen decoy and through effects on Fc receptor-bearing leukocytes. *J Virol* 2008;**82**:12191–204.

Bull RA, Eden J-S, Luciani F *et al.* Contribution of Intra- and Interhost Dynamics to Norovirus Evolution. *J Virol* 2012;**86**:3219–29.

Calderón A, Pozo F, Calvo C *et al.* Genetic variability of respiratory syncytial virus A in hospitalized children in the last five consecutive winter seasons in Central Spain. *J Med Virol* 2017;**89**:767–74.

Canchaya C, Proux C, Fournous G *et al.* Prophage genomics. *Microbiol Mol Biol Rev* 2003;**67**:238–76, table of contents.

Cane PA. Analysis of linear epitopes recognised by the primary human antibody response to a variable region of the attachment (G) protein of respiratory syncytial virus. *J Med Virol* 1997;**51**:297–304.

Cane PA. Molecular epidemiology of respiratory syncytial virus. *Rev Med Virol* 2001;**11**:103–16.

Cane PA, Matthews DA, Pringle CR. Identification of variable domains of the attachment (G) protein of subgroup A respiratory syncytial viruses. *J Gen Virol* 1991;**72**:2091–6.

Cane PA, Pringle CR. Evolution of subgroup A respiratory syncytial virus: evidence

for progressive accumulation of amino acid changes in the attachment protein. *J Virol* 1995;**69**:2918–25.

Centers for Disease Control and Prevention (CDC). Respiratory syncytial virus activity--United States, 2003-2004. *MMWR Morb Mortal Wkly Rep* 2004;**53**:1159–60.

Chanock R, Finberg L. Recovery from infants with respiratory illness of a virus related to chimpanzee coryza agent (CCA). II. Epidemiologic aspects of infection in infants and young children. *Am J Hyg* 1957;**66**:291–300.

Chanock R, Roizman B, Myers R. Recovery from infants with respiratory illness of a virus related to chimpanzee coryza agent (CCA). I. Isolation, properties and characterization. *Am J Hyg* 1957;**66**:281–90.

Chernomor O, von Haeseler A, Minh BQ. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol* 2016:syw037.

Cherrie AH, Anderson K, Wertz GW *et al.* Human cytotoxic T cells stimulated by antigen on dendritic cells recognize the N, SH, F, M, 22K, and 1b proteins of respiratory syncytial virus. *J Virol* 1992;**66**:2102–10.

Chirkova T, Boyoglu-Barnum S, Gaston KA *et al.* Respiratory Syncytial Virus G Protein CX3C Motif Impairs Human Airway Epithelial and Immune Cell Responses. *J Virol* 2013;**87**:13466–79.

Choudhary ML, Wadhwa BS, Jadhav SM *et al.* Complete genom sequences of two human respiratory scyntival virus genome A strains from India, RSV-A?NIV1114046/11 and RSV-A/NIV1114073/11. *Genome Announc* 2013;**1**:e00165-13.

Cobbin JCA, Alfelali M, Barasheed O *et al.* Multiple Sources of Genetic Diversity of Influenza A Viruses during the Hajj. *J Virol* 2017;**91**:e00096-17.

Collins PL, Melero JA. Progress in understanding and controlling respiratory syncytial virus: Still crazy after all these years. *Virus Res* 2011;**162**:80–99.

Comas-García A, Noyola DE, Cadena-Mota S *et al.* Respiratory Syncytial Virus-A ON1 Genotype Emergence in Central Mexico in 2009 and Evidence of Multiple Duplication Events. *J Infect Dis* 2018;**217**:1089–98.

Conenello GM, Zamarin D, Perrone LA *et al.* A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence. *PLoS Pathog* 2007;**3**:1414–21.

Connors M, Collins PL, Firestone CY *et al.* Respiratory syncytial virus (RSV) F, G, M2 (22K), and N proteins each induce resistance to RSV challenge, but resistance induced by M2 and N proteins is relatively short-lived. *J Virol* 1991;**65**:1634–7.

Connors M, Crowe JE, Firestone CY *et al.* A Cold-Passaged, Attenuated Strain of Human Respiratory Syncytial Virus Contains Mutations in the F and L Genes. *Virology* 1995;**208**:478–84.

Cote PJ, Fernie BF, Ford EC *et al.* Monoclonal antibodies to respiratory syncytial virus: Detection of virus neutralization and other antigen-antibody systems using infected human and murine cells. *J Virol Methods* 1981;**3**:137–47.

Cristina J, López JA, Albó C *et al.* Analysis of genetic variability in human respiratory syncytial virus by the RNase a mismatch cleavage method: Subtype divergence and heterogeneity. *Virology* 1990;**174**:126–34.

Dapat IC, Shobugawa Y, Sano Y *et al.* New genotypes within respiratory syncytial virus group B genotype BA in Niigata, Japan. *J Clin Microbiol* 2010;**48**:3423–7.

Deurenberg RH, Bathoorn E, Chlebowicz MA *et al.* Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol*

2017;**243**:16–24.

Domingo E, Diez J, Martinez MA *et al.* New observations on antigenic diversification of RNA viruses: Antigenic variation is not dependent on immune selection. *J Gen Virol* 1993;**74**:2039–45.

Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. *Microbiol Mol Biol Rev* 2012;**76**:159–216.

Drummond AJ, Ho SYW, Phillips MJ *et al.* Relaxed phylogenetics and dating with confidence. *PLoS Biol* 2006;**4**:699–710.

Drummond AJ, Rambaut A, Shapiro B *et al.* Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005;**22**:1185–92.

Drummond AJ, Suchard MA, Xie D *et al.* Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012;**29**:1969–73.

Dudas G, Carvalho LM, Bedford T *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* 2017;**544**:309–15.

Duvvuri VR, Granados A, Rosenfeld P *et al.* Genetic diversity and evolutionary insights of respiratory syncytial virus A ON1 genotype: global and local transmission dynamics. *Sci Rep* 2015;**5**:14268.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.

Edwards CJ, Suchard MA, Lemey P *et al.* Ancient hybridization and an irish origin for the modern polar bear matriline. *Curr Biol* 2011, DOI: 10.1016/j.cub.2011.05.058.

Eiter T, Faber W, Fink M *et al.* Complexity of answer set checking and bounded predicate arities for non-ground Answer Set Programming. *CEUR Workshop*

Proc 2003;**78**:69–83.

Emukule GO, Khagayi S, McMorrow ML *et al.* The burden of influenza and rsv among inpatients and outpatients in rural western Kenya, 2009-2012. *PLoS One* 2014;**9**:e105543.

Eshaghi AR, Duvvuri VR, Lai R *et al.* Genetic variability of human respiratory syncytial virus a strains circulating in Ontario: A novel genotype with a 72 nucleotide G gene duplication. *PLoS One* 2012;**7**:e32807.

Espinoza JA, Bohmwald K, Céspedes PF *et al.* Modulation of host adaptive immunity by hRSV proteins. *Virulence* 2014;**5**:740–51.

Ewels P, Magnusson M, Lundin S *et al.* MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.

Fall A, Dia N, Cisse EHAK *et al.* Epidemiology and Molecular Characterization of Human Respiratory Syncytial Virus in Senegal after Four Consecutive Years of Surveillance, 2012–2015. Schildgen O (ed.). *PLoS One* 2016;**11**:e0157163.

Falsey A. Respiratory Syncytial Virus Infection in Adults. *Semin Respir Crit Care Med* 2007;**28**:171–81.

Falsey AR, Hennessey PA, Formica MA *et al.* Respiratory syncytial virus infection in elderly and high-risk adults. *N Engl J Med* 2005;**352**:1749–59.

Faria NR, Quick J, Claro IM *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 2017;**546**:406–10.

Faria NR, Suchard MA, Rambaut A *et al.* Toward a quantitative understanding of viral phylogeography. *Curr Opin Virol* 2011;**1**:423–9.

Faria NR, Suchard MA, Rambaut A *et al.* Simultaneously reconstructing viral crossspecies transmission history and identifying the underlying constraints. *Philos Trans R Soc B Biol Sci* 2013;**368**, DOI: 10.1098/rstb.2012.0196.

Fauci AS. Infectious Diseases: Considerations for the 21st Century. *Clin Infect Dis* 2001;**32**:675–85.

Feikin DR, Olack B, Bigogo GM *et al.* The burden of common infectious disease syndromes at the clinic and household level from population-based surveillance in rural and Urban Kenya. *PLoS One* 2011;**6**:e16085.

Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. *Can J Stat* 2008;**36**:355–68.

Fodha I, Vabret A, Trabelsi A *et al.* Epidemiological and Antigenic Analysis of Respiratory Syncytial Virus in Hospitalised Tunisian Children, from 2000 to 2002. *J Med Virol* 2004;**72**:683–7.

Folkerts G, Busse WW, Nijkamp FP *et al.* Virus-induced airway hyperresponsiveness and asthma. *Am J Respir Crit Care Med* 1998;**157**:1708–20.

Fuentes S, Coyle EM, Beeler J *et al.* Antigenic Fingerprinting following Primary RSV Infection in Young Children Identifies Novel Antigenic Sites and Reveals Unlinked Evolution of Human Antibody Repertoires to Fusion and Attachment Glycoproteins. Wilson PC (ed.). *PLoS Pathog* 2016;**12**:e1005554.

Fuentes S, Tran KC, Luthra P *et al.* Function of the Respiratory Syncytial Virus Small Hydrophobic Protein. *J Virol* 2007;**81**:8361–6.

Garcia-barreno B, Palomo C, Penas C *et al.* Marked Differences in the Antigenic Structure of Human Respiratory Syncytial Virus F and G Glycoproteins. 1989;**63**:925–32.

García O, Martín M, Dopazo J *et al.* Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G glycoprotein. *J Virol* 1994;**68**:5448–59.

Garofalo RP, Patti J, Hintz KA *et al.* Macrophage Inflammatory Protein–1α (Not T

Helper Type 2 Cytokines) Is Associated with Severe Forms of Respiratory Syncytial Virus Bronchiolitis. *J Infect Dis* 2001;**184**:393–9.

Gaymard A, Bouscambert-Duchamp M, Pichon M *et al.* Genetic characterization of respiratory syncytial virus highlights a new BA genotype and emergence of the ON1 genotype in Lyon, France, between 2010 and 2014. *J Clin Virol* 2018;**102**:12–8.

Githinji G, Agoti CN, Kibinge N *et al.* Assessing the utility of minority variant composition in elucidating RSV transmission pathways. *bioRxiv* 2018:411512.

Glezen WP, Taber LH, Frank AL *et al.* Risk of primary infection and reinfection with respiratory syncytial virus. *Am J Dis Child* 1986;**140**:543–6.

Goddard NL, Cooke MC, Gupta RK *et al.* Timing of monoclonal antibody for seasonal RSV prophylaxis in the United Kingdom. *Epidemiol Infect* 2007;**135**:159–62.

Gostin LO. Our shared vulnerability to dangerous pathogens. *Med Law Rev* 2017;**25**:185–99.

Goya S, Valinotto LE, Tittarelli E *et al.* An optimized methodology for whole genome sequencing of RNA respiratory viruses from nasopharyngeal aspirates. *PLoS One* 2018;**13**, DOI: 10.1371/journal.pone.0199714.

Grad YH, Newman R, Zody M *et al.* Within-Host Whole-Genome Deep Sequencing and Diversity Analysis of Human Respiratory Syncytial Virus Infection Reveals Dynamics of Genomic Diversity in the Absence and Presence of Immune Pressure. *J Virol* 2014;**88**:7286–93.

Grenfell BT, Pybus OG, Gog JR *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004;**303**:327–32.

Grubaugh ND, Andersen KG. Experimental Evolution to Study Virus Emergence.

Cell 2017;**169**:1–3.

Grubaugh ND, Ladner JT, Kraemer MUG *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* 2017;**546**:401–5.

Gunson RN, Collins TC, Carman WF. Real-time RT-PCR detection of 12 respiratory viral infections in four triplex reactions. *J Clin Virol* 2005;**33**:341–4.

Gupta R, Jung E, Brunak S. NetNGlyc: Prediction of N-glycosylation sites in human proteins. *Prep* 2004;**46**:2004.

Ha Do LA, Wilm A, Van Doorn HR *et al.* Direct whole-genome deep-sequencing of human respiratory syncytial virus A and B from Vietnamese children identifies distinct patterns of inter- and intra-host evolution. *J Gen Virol* 2015;**96**:3470–83.

Habibi MS, Jozwik A, Makris S *et al.* Impaired Antibody-mediated Protection and Defective IgA B-Cell Memory in Experimental Infection of Adults with Respiratory Syncytial Virus. *Am J Respir Crit Care Med* 2015;**191**:1040–9.

Hall CB, Walsh EE, Long CE *et al.* Immunity to and frequency of reinfection with respiratory syncytial virus. *J Infect Dis* 1991;**163**:693–8.

Hall CB, Walsh EE, Schnabel KC *et al.* Occurrence of groups A and B of respiratory syncytial virus over 15 years: Associated epidemiologic and clinical characteristics in hospitalized and ambulatory children. *J Infect Dis* 1990;**162**:1283–90.

Hall T. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 1999;**41**:95–8.

Hammitt LL, Kazungu S, Morpeth SC *et al.* A Preliminary Study of Pneumonia Etiology Among Hospitalized Children in Kenya. *Clin Infect Dis* 2012;**54**:S190–9.

Hammitt LL, Kazungu S, Welch S *et al.* Added Value of an Oropharyngeal Swab in Detection of Viruses in Children Hospitalized with Lower Respiratory Tract Infection. *J Clin Microbiol* 2011;**49**:2318–20.

Harmon SB, Wertz GW. Transcriptional termination modulated by nucleotides outside the characterized gene end sequence of respiratory syncytial virus. *Virology* 2002;**300**:304–15.

Hatta M, Gao P, Halfmann P *et al.* Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* 2001;**293**:1840–2.

Hause AM, Henke DM, Avadhanula V *et al.* Sequence variability of the respiratory syncytial virus (RSV) fusion gene among contemporary and historical genotypes of RSV/A and RSV/B. *PLoS One* 2017;**12**, DOI: 10.1371/journal.pone.0175792.

Haynes AK, Manangan AP, Iwane MK *et al.* Respiratory syncytial virus circulation in seven countries with global disease detection regional centers. *J Infect Dis* 2013;**208**:S246-54.

Henderson FW, Collier AM, Clyde WA *et al.* Respiratory-syncytial-virus infections, reinfections and immunity. A prospective, longitudinal study in young children. *N Engl J Med* 1979;**300**:530–4.

Hendricks DA, Baradaran K, McIntosh K *et al.* Appearance of a soluble form of the G protein of respiratory syncytial virus in fluids of infected cells. *J Gen Virol* 1987;**68**:1705–14.

Hendry RM, Burns JC, Walsh EE *et al.* Strain-Specific Serum Antibody Responses in Infants Undergoing Primary Infection with Respiratory Syncytial Virus. *J Infect Dis* 1988;**157**:640–7.

Hendry RM, Pierik LT, McIntosh K. Prevalence of Respiratory Syncytial Virus Subgroups Over Six Consecutive Outbreaks: 1981-1987. *J Infect Dis*

1989;**160**:185–90.

Henn MR, Boutwell CL, Charlebois P *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012;**8**, DOI: 10.1371/journal.ppat.1002529.

Holland J, Spindler K, Horodyski F *et al.* Rapid evolution of RNA genomes. *Science* 1982;**215**:1577–85.

Holmes EC. Evolutionary History and Phylogeography of Human Viruses. *Annu Rev Microbiol* 2008;**62**:307–28.

Holmes EC, Dudas G, Rambaut A *et al.* The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* 2016;**538**:193–200.

Hotard AL, Laikhter E, Brooks K *et al.* Functional Analysis of the 60 Nucleotide Duplication in the Respiratory Syncytial Virus Buenos Aires Strain Attachment Glycoprotein. *J Virol* 2015;**89**:JVI.01045-15.

Hunt M, Newbold C, Berriman M *et al.* A comprehensive evaluation of assembly scaffolding tools. *Genome Biol* 2014, DOI: 10.1186/gb-2014-15-3-r42.

Ikonen N, Savolainen-Kopra C, Enstone JE *et al.* Deposition of respiratory virus pathogens on frequently touched surfaces at airports. *BMC Infect Dis* 2018;**18**, DOI: 10.1186/s12879-018-3150-5.

J. Spielman S. phyphy: Python package for facilitating the execution and parsing of HyPhy standard analyses. *J Open Source Softw* 2018;**3**:514.

Johnson PR, Collins PL. The A and B subgroups of human respiratory syncytial virus: Comparison of intergenic and gene-overlap sequences. *J Gen Virol* 1988;**69**:2901–6.

Johnson PR, Collins PL. Sequence comparison of the phosphoprotein mRNAs of antigenic subgroups A and B of human respiratory syncytial virus identifies a

highly divergent domain in the predicted protein. *J Gen Virol* 1990;**71**:481–5.

Johnson PR, Spriggs MK, Olmsted RA *et al.* The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. *Proc Natl Acad Sci U S A* 1987;**84**:5625–9.

Johnson SM, McNally BA, Ioannidis I *et al.* Respiratory Syncytial Virus Uses CX3CR1 as a Receptor on Primary Human Airway Epithelial Cultures. *PLoS Pathog* 2015, DOI: 10.1371/journal.ppat.1005318.

Jones KE, Patel NG, Levy MA *et al.* Global trends in emerging infectious diseases. *Nature* 2008;**451**:990–3.

Jozwik A, Habibi MS, Paras A *et al.* RSV-specific airway resident memory CD8+ T cells and differential disease severity after experimental human infection. *Nat Commun* 2015, DOI: 10.1038/ncomms10224.

Kamoun EA, Youssef ME, Abu-Saied MA *et al.* Ion conducting nanocomposite membranes based on PVA-HA-HAP for fuel cell application: II. Effect of modifier agent of PVA on membrane properties. *Int J Electrochem Sci* 2015;**10**:6627–44.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–80.

Katz MA, Lebo E, Emukule G *et al.* Epidemiology, seasonality, and burden of influenza and influenza-like illness in Urban and Rural Kenya, 2007-2010. *J Infect Dis* 2012;**206**:S53-60.

Katz MA, Muthoka P, Emukule GO *et al.* Results from the first six years of national sentinel surveillance for influenza in Kenya, July 2007-June 2013. *PLoS One* 2014;**9**:e98615.

Katzov-Eckert H, Botosso VF, Neto EA *et al.* Phylodynamics and Dispersal of HRSV Entails Its Permanence in the General Population in between Yearly Outbreaks in Children. *PLoS One* 2012;**7**:e41953.

Kearse M, Moir R, Wilson A *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;**28**:1647–9.

Kenya National Bureau of Statistics. County Statistics. *Cty Stat* 2013.

Kenya National Bureau of Statistics (KNBS), Society for International Development – East Africa (SID). *Exploring Kenya's Inequality Pulling Apart or Pooling Together?: National Report*. Nairobi, Kenya, 2013.

Khor CS, Sam IC, Hooi PS *et al.* Displacement of predominant respiratory syncytial virus genotypes in Malaysia between 1989 and 2011. *Infect Genet Evol* 2013;**14**:357–60.

Kim YJ, Kim DW, Lee WJ *et al.* Rapid replacement of human respiratory syncytial virus A with the ON1 genotype having 72 nucleotide duplication in G gene. *Infect Genet Evol* 2014;**26**:103–12.

Kinyanjui TM, House TA, Kiti MC *et al.* Vaccine induced herd immunity for control of respiratory syncytial virus disease in a low-income country setting. *PLoS One* 2015;**10**, DOI: 10.1371/journal.pone.0138018.

Kiti MC, Kinyanjui TM, Koech DC *et al.* Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya. *PLoS One* 2014;**9**, DOI: 10.1371/journal.pone.0104786.

Kiyuka PK, Agoti CN, Munywoki PK *et al.* Human coronavirus NL63 molecular epidemiology and evolutionary patterns in rural coastal Kenya. *J Infect Dis* 2018;**217**:1728–39.

Komissarova N, Kashlev M. Transcriptional arrest: Escherichia coli RNA polymerase translocates backward, leaving the 3' end of the RNA intact and extruded. *Proc Natl Acad Sci USA* 1997;**94**:1755–60.

Korsun N, Angelova S, Tzotcheva I *et al.* Prevalence and genetic characterisation of respiratory syncytial viruses circulating in Bulgaria during the 2014/15 and 2015/16 winter seasons. *Pathog Glob Health* 2017, DOI: 10.1080/20477724.2017.1375708.

Kosakovsky Pond SL, Frost SDW, Pond SLK *et al.* Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Mol Biol Evol* 2005;**22**:1208–22.

Krzyzaniak MA, Zumstein MT, Gerez JA *et al.* Host Cell Entry of Respiratory Syncytial Virus Involves Macropinocytosis Followed by Proteolytic Activation of the F Protein. Pekosz A (ed.). *PLoS Pathog* 2013;**9**:e1003309.

Kumaria R, Iyer L, Hibberd ML *et al.* Whole genome characterization of non-tissue culture adapted HRSV strains in severely infected children. *Virol J* 2011;**8**:372.

Kuo L, Fearns R, Collins PL. Analysis of the gene start and gene end signals of human respiratory syncytial virus: quasi-templated initiation at position 1 of the encoded mRNA. *J Virol* 1997;**71**:4944–53.

Lam TTY, Wang J, Shen Y *et al.* The genesis and source of the H7N9 influenza viruses causing human infections in China. *Nature* 2013;**502**:241–4.

Langmead B, Trapnell C, Pop M *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.

Larsson A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 2014;**30**:3276–8.

Lê S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. *J Stat*

Softw 2008;**25**:253–8.

Lee W-J, Kim Y -j., Kim D-W *et al.* Complete Genome Sequence of Human Respiratory Syncytial Virus Genotype A with a 72-Nucleotide Duplication in the Attachment Protein G Gene. *J Virol* 2012;**86**:13810–1.

Lemey P, Rambaut A, Bedford T *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. Ferguson NM (ed.). *PLoS Pathog* 2014;**10**:e1003932.

Lemey P, Rambaut A, Drummond AJ *et al.* Bayesian phylogeography finds its roots. *PLoS Comput Biol* 2009;**5**, DOI: 10.1371/journal.pcbi.1000520.

Lemey P, Rambaut A, Welch JJ *et al.* Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol* 2010;**27**:1877–85.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.

Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

López JA, Bustos R, Orvell C *et al.* Antigenic structure of human respiratory syncytial virus fusion glycoprotein. *J Virol* 1998;**72**:6922–8.

López JA, Peñas C, García-Barreno B *et al.* Location of a highly conserved neutralizing epitope in the F glycoprotein of human respiratory syncytial virus. *J Virol* 1990;**64**:927–30.

Madhi SA, Schoub B, Simmank K *et al.* Increased burden of respiratory viral associated severe lower respiratory tract infections in children infected with human immunodeficiency virus type-1. *J Pediatr* 2000;**137**:78–84.

Magoc T, Pabinger S, Canzar S *et al.* GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 2013;**29**:1718–25.

Martinello RA, Chen MD, Weibel C *et al.* Correlation between Respiratory Syncytial Virus Genotype and Severity of Illness. *J Infect Dis* 2002;**186**:839–42.

Martínez I, Dopazo J, Melero JA. Antigenic structure of the human respiratory syncytial virus G glycoprotein and relevance of hypermutation events for the generation of antigenic variants. *J Gen Virol* 1997;**78**:2419–29.

McLellan JS, Chen M, Leung S *et al.* Structure of RSV Fusion Glycoprotein Trimer Bound to a Prefusion-Specific Neutralizing Antibody. *Science (80-)* 2013;**340**:1113–7.

Meerhoff TJ, Paget JW, Kimpen JL *et al.* Variation of respiratory syncytial virus and the relation with meteorological factors in different winter seasons. *Pediatr Infect Dis J* 2009;**28**:860–6.

Melero JA, Garcia-Barreno B, Martinez I *et al.* Antigenic structure, evolution and immunobiology of human respiratory syncytial virus attachment (G) protein. *J Gen Virol* 1997;**78**:2411–8.

Melero JA, Moore ML. Influence of Respiratory Syncytial Virus Strain Differences on Pathogenesis and Immunity. *Current Topics in Microbiology and Immunology*. Vol 372. NIH Public Access, 2013, 59–82.

Metsky HC, Matranga CB, Wohl S *et al.* Zika virus evolution and spread in the Americas. *Nature* 2017;**546**:411–5.

Metzker ML. Sequencing technologies the next generation. *Nat Rev Genet* 2010;**11**:31–46.

Michael Hendry R, Talis AL, Godfrey E *et al.* Concurrent circulation of antigenically distinct strains of respiratory syncytial virus during community outbreaks. *J Infect Dis* 1986;**153**:291–7.

Minin VN, Suchard M. Counting labeled transitions in continuous-time Markov

models of evolution. *J Math Biol* 2008, DOI: 10.1007/s00285-007-0120-8.

Mitnaul LJ, Matrosovich MN, Castrucci MR *et al.* Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *J Virol* 2000;**74**:6015–20.

Mlinaric-Galinovic G, Tabain I, Kukovec T *et al.* Analysis of biennial outbreak pattern of respiratory syncytial virus according to subtype (A and B) in the Zagreb region. *Pediatr Int* 2012;**54**:331–5.

Moïsi JC, Nokes DJ, Gatakaa H *et al.* Sensitivity of hospital-based surveillance for severe disease: a geographic information system analysis of access to care in Kilifi district, Kenya. *Bull World Health Organ* 2011, DOI: 10.2471/BLT.10.080796.

Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. *Nature* 2004;**430**:242–9.

Morey M, Fernández-Marmiesse A, Castiñeiras D *et al.* A glimpse into past, present, and future DNA sequencing. *Mol Genet Metab* 2013;**110**:3–24.

Morris JA, Blount RE, Savage RE. Recovery of Cytopathogenic Agent from Chimpanzees with Goryza. *Exp Biol Med* 1956;**92**:544–9.

Moudy RM, Sullender WM, Wertz GW. Variations in intergenic region sequences of Human respiratory syncytial virus clinical isolates: Analysis of effects on transcriptional regulation. *Virology* 2004;**327**:121–33.

Moura FE, Nunes IF, Silva Jr. GB *et al.* Respiratory syncytial virus infections in northeastern Brazil: seasonal trends and general aspects. *Am J Trop Med Hyg* 2006;**74**:165–7.

Moura FEA, Blanc A, Frabasile S *et al.* Genetic diversity of respiratory syncytial virus isolated during an epidemic period from children of Northeastern Brazil. *J*

Med Virol 2004;**74**:156–60.

Mufson MA, Orvell C, Rafnar B *et al.* Two Distinct Subtypes of Human Respiratory Syncytial Virus. *J Veneral Virol* 1985;**66 (Pt 10**:2111–24.

Munywoki PK, Hamid F, Mutunga M *et al.* Improved detection of respiratory viruses in pediatric outpatients with acute respiratory illness by real-time PCR using nasopharyngeal flocked swabs. *J Clin Microbiol* 2011;**49**:3365–7.

Munywoki PK, Koech DC, Agoti CN *et al.* The source of respiratory syncytial virus infection in infants: A household cohort study in rural Kenya. *J Infect Dis* 2014;**209**:1685–92.

Murray EL, Klein M, Brondi L *et al.* Rainfall, household crowding, and acute respiratory infections in the tropics. *Epidemiol Infect* 2012;**140**:78–86.

Murrell B, Moola S, Mabona A *et al.* FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Mol Biol Evol* 2013;**30**:1196–205.

Murrell B, Weaver S, Smith MD *et al.* Gene-wide identification of episodic selection. *Mol Biol Evol* 2015;**32**:1365–71.

Murrell B, Wertheim JO, Moola S *et al.* Detecting individual sites subject to episodic diversifying selection. Malik HS (ed.). *PLoS Genet* 2012;**8**:e1002764.

Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet* 2013, DOI: 10.1038/nrg3367.

Nair H, Nokes DJ, Gessner BD *et al.* Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet (London, England)* 2010;**375**:1545–55.

Neuzil KM. Progress toward a Respiratory Syncytial Virus Vaccine. *Clin Vaccine Immunol* 2016, DOI: 10.1128/CVI.00037-16.

Newman AP, Reisdorf E, Beinemann J *et al.* Human case of swine influenza A

(H1N1) triple reassortant virus infection, Wisconsin. *Emerg Infect Dis* 2008;**14**:1470–2.

van Niekerk S, Venter M. Replacement of Previously Circulating Respiratory Syncytial Virus Subtype B Strains with the BA Genotype in South Africa. *J Virol* 2011;**85**:8789–97.

Nokes DJ. Respiratory syncytial virus disease burden in the developing world. In: Cane P (ed.). *Perspectives in Medical Virology*. Vol 14. Elsevier, 2007, 183–232.

Nokes DJ, Ngama M, Bett A *et al.* Incidence and Severity of Respiratory Syncytial Virus Pneumonia in Rural Kenyan Children Identified through Hospital Surveillance. *Clin Infect Dis* 2009;**49**:1341–9.

Nokes DJ, Okiro EA, Ngama M *et al.* Respiratory Syncytial Virus Epidemiology in a Birth Cohort from Kilifi District, Kenya: Infection during the First Year of Life. *J Infect Dis* 2004;**190**:1828–32.

Nokes DJ, Okiro EA, Ngama M *et al.* Respiratory Syncytial Virus Infection and Disease in Infants and Young Children Observed from Birth in Kilifi District, Kenya. *Clin Infect Dis* 2008;**46**:50–7.

Nyiro JU, Kombe IK, Sande CJ *et al.* Defining the vaccination window for respiratory syncytial virus (RSV) using age-seroprevalence data for children in Kilifi, Kenya. Tregoning JS (ed.). *PLoS One* 2017;**12**:e0177803.

Nyiro JU, Munywoki P, Kamau E *et al.* Surveillance of respiratory viruses in the outpatient setting in rural coastal Kenya: baseline epidemiological observations. *Wellcome Open Res* 2018;**3**:89.

Obando-Pacheco P, Justicia-Grande AJ, Rivero-Calle I *et al.* Respiratory syncytial virus seasonality: A global overview. *J Infect Dis* 2018;**217**:1356–64.

Odhiambo FO, Laserson KF, Sewe M *et al.* Profile: The KEMRI/CDC health and

demographic surveillance system-Western Kenya. *Int J Epidemiol* 2012;**41**:977–87.

Olmsted RA, Elango N, Prince GA *et al.* Expression of the F glycoprotein of respiratory syncytial virus by a recombinant vaccinia virus: comparison of the individual contributions of the F and G glycoproteins to host immunity. *Proc Natl Acad Sci U S A* 1986;**83**:7462–6.

Onyango CO, Njeru R, Kazungu S *et al.* Influenza Surveillance Among Children With Pneumonia Admitted to a District Hospital in Coastal Kenya, 2007-2010. *J Infect Dis* 2012a;**206**:S61–7.

Onyango CO, Welch SR, Munywoki PK *et al.* Molecular epidemiology of human rhinovirus infections in Kilifi, coastal Kenya. *J Med Virol* 2012b;**84**:823–31.

Openshaw PJM. Potential therapeutic implications of new insights into respiratory syncytial virus disease. *Respir Res* 2002, DOI: 10.1186/rr184.

Otieno JR, Agoti CN, Gitahi CW *et al.* Molecular Evolutionary Dynamics of Respiratory Syncytial Virus Group A in Recurrent Epidemics in Coastal Kenya. García-Sastre A (ed.). *J Virol* 2016;**90**:4990–5002.

Otieno JR, Kamau EM, Agoti CN *et al.* Spread and evolution of respiratory syncytial virus a genotype ON1, coastal Kenya, 2010-2015. *Emerg Infect Dis* 2017;**23**:264–71.

Otieno JR, Kamau EM, Oketch JW *et al.* Erratum: Whole genome analysis of local Kenyan and global sequences unravels the epidemiological and molecular evolutionary dynamics of RSV genotype ON1 strains. *Virus Evol* 2018;**4**:vey036.

Out of Africa. *Nature* 2014;**514**:139–139.

Owor BE, Masankwa GN, Mwango LC *et al.* Human metapneumovirus

epidemiological and evolutionary patterns in Coastal Kenya, 2007-11. *BMC Infect Dis* 2016;**16**:301.

Panayiotou C, Richter J, Koliou M *et al.* Epidemiology of respiratory syncytial virus in children in Cyprus during three consecutive winter seasons (2010-2013): Age distribution, seasonality and association between prevalent genotypes and disease severity. *Epidemiol Infect* 2014;**142**:2406–11.

Park DJ, Dudas G, Wohl S *et al.* Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell* 2015;**161**:1516–26.

Park DJ, Tomkins-Tinch C, Ye S *et al.* Broad Institute viral-ngs. 2016.

Park E, Park PH, Huh JW *et al.* Molecular and clinical characterization of human respiratory syncytial virus in South Korea between 2009 and 2014. *Epidemiol Infect* 2017, DOI: 10.1017/S0950268817002230.

Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect Genet Evol* 2008;**8**:239–46.

Pastor-Satorras R, Castellano C, Van Mieghem P *et al.* Epidemic processes in complex networks. *Rev Mod Phys* 2015;**87**, DOI: 10.1103/RevModPhys.87.925.

PATH. RSV Vaccine and mAb Snapshot. *https://path.org/resources/rsv-vaccine-and-mab-snapshot* 2019.

Peck KM, Lauring AS. Complexities of Viral Mutation Rates. *J Virol* 2018, DOI: 10.1128/JVI.01031-17.

Pereira AC, Monteiro SN, Assis FS *et al.* Charpy Toughness Behavior of Fique Fabric Reinforced Polyester Matrix Composites. *Miner Met Mater Ser* 2017;**Part F7**:3–9.

Peret TC, Hall CB, Hammond GW *et al.* Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America.

J Infect Dis 2000;**181**:1891–6.

Peret TC, Hall CB, Schnabel KC *et al.* Circulation patterns of genetically distinct group A and B strains of human respiratory syncytial virus in a community. *J Gen Virol* 1998;**79 (Pt 9)**:2221–9.

Pierangeli A, Trotta D, Scagnolari C. Rapid spread of the novel respiratory syncytial virus A ON1 genotype, central Italy, 2011 to 2013. *Euro Surveill* 2014;**19**:pii: 20843.

Piñana M, Vila J, Gimferrer L *et al.* Novel human metapneumovirus with a 180-nucleotide duplication in the G gene. *Future Microbiol* 2017;**12**:565–71.

Polack FP, Irusta PM, Hoffman SJ *et al.* The cysteine-rich region of respiratory syncytial virus attachment protein inhibits innate immunity elicited by the virus and endotoxin. *Proc Natl Acad Sci* 2005;**102**:8996–9001.

Pond SLK, Frost SDW, Muse S V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005;**21**:676–9.

Preidis GA, McCollum ED, Mwansambo C *et al.* Pneumonia and Malnutrition are Highly Predictive of Mortality among African Children Hospitalized with Human Immunodeficiency Virus Infection or Exposure in the Era of Antiretroviral Therapy. *J Pediatr* 2011;**159**:484–9.

Pretorius MA, Van Niekerk S, Tempia S *et al.* Replacement and positive evolution of subtype A and B respiratory syncytial virus G-protein genotypes from 1997-2012 in South Africa. *J Infect Dis* 2013;**208**:S227-37.

Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**, DOI: 10.1371/journal.pone.0009490.

Prifert C, Streng A, Krempl CD *et al.* Novel respiratory syncytial virus a genotype,

Germany, 2011-2012. *Emerg Infect Dis* 2013;**19**:1029–30.

Pybus O. Taming The Beast: Introduction to infectious disease phylodynamics. 2016:4.

Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* 2009;**10**:540–50.

Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.

Raghwani J, Thompson RN, Koelle K. Selection on non-antigenic gene segments of seasonal influenza A virus and its impact on adaptive evolution. *Virus Evol* 2017;**3**, DOI: 10.1093/ve/vex034.

Rajala MS, Sullender WM, Prasad a K *et al.* Genetic Variability Among Group A and B Respiratory Syncytial Virus Isolates From a Large Referral Hospital in New Delhi , India. *Society* 2003;**41**:2311–6.

Rambaut A, Lam TT, Max Carvalho L *et al.* Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016;**2**:vew007.

Rebuffo-Scheer C, Bose M, He J *et al.* Whole genome sequencing and evolutionary analysis of human respiratory syncytial virus A and B from Milwaukee, WI 1998-2010. Lopez-Galindez C (ed.). *PLoS One* 2011;**6**:e25468.

Ren L, Xia Q, Xiao Q *et al.* The genetic variability of glycoproteins among respiratory syncytial virus subtype A in China between 2009 and 2013. *Infect Genet Evol* 2014;**27**:339–47.

Roberts SR, Lichtenstein D, Ball LA *et al.* The membrane-associated and secreted forms of the respiratory syncytial virus attachment glycoprotein G are synthesized from alternative initiation codons. *J Virol* 1994;**68**:4538–46.

Rodrigo AG. Coalescent estimates of HIV-1 generation time in vivo. *Proc Natl Acad Sci* 1999, DOI: 10.1073/pnas.96.5.2187.

Rodriguez-Roche R, Blanc H, Bordería A V. *et al.* Increasing Clinical Severity during a Dengue Virus Type 3 Cuban Epidemic: Deep Sequencing of Evolving Viral Populations. *J Virol* 2016;**90**:4320–33.

Rojo GL, Goya S, Orellana M *et al.* Unravelling respiratory syncytial virus outbreaks in Buenos Aires, Argentina: Molecular basis of the spatio-temporal transmission. *Virology* 2017;**508**:118–26.

RStudio Team -. RStudio: Integrated Development for R. *[Online] RStudio, Inc, Boston, MA URL http//www rstudio com* 2016:RStudio, Inc., Boston, MA.

Russell CA, Jones TC, Barr IG *et al.* The global circulation of seasonal influenza A (H3N2) viruses. *Science* 2008;**320**:340–6.

Russell RF, McDonald JU, Ivanova M *et al.* Partial Attenuation of Respiratory Syncytial Virus with a Deletion of a Small Hydrophobic Gene Is Associated with Elevated Interleukin-1β Responses. Lyles DS (ed.). *J Virol* 2015;**89**:8974–81.

Saikusa M, Kawakami C, Nao N *et al.* 180-Nucleotide Duplication in the G Gene of Human metapneumovirus A2b Subgroup Strains Circulating in Yokohama City, Japan, since 2014. *Front Microbiol* 2017a;**8**, DOI: 10.3389/fmicb.2017.00402.

Saikusa M, Nao N, Kawakami C *et al.* A novel 111-nucleotide duplication in the G gene of human metapneumovirus. *Microbiol Immunol* 2017b;**61**:507–12.

Sande CJ, Mutunga MN, Medley GF *et al.* Group-and genotype-specific neutralizing antibody responses against respiratory syncytial virus in infants and young children with severe pneumonia. *J Infect Dis* 2013;**207**:489–92.

Schobel SA, Stucker KM, Moore ML *et al.* Respiratory Syncytial Virus whole-genome sequencing identifies convergent evolution of sequence duplication in

the C-terminus of the G gene. *Sci Rep* 2016;**6**, DOI: 10.1038/srep26311.

Scott JAG, Bauni E, Moisi JC *et al.* Profile: The Kilifi health and demographic surveillance system (KHDSS). *Int J Epidemiol* 2012;**41**:650–7.

Scott PD, Ochola R, Ngama M *et al.* Molecular epidemiology of respiratory syncytial virus in Kilifi District, Kenya. *J Med Virol* 2004;**74**:344–54.

Scott PD, Ochola R, Ngama M *et al.* Molecular Analysis of Respiratory Syncytial Virus Reinfections in Infants from Coastal Kenya. *J Infect Dis* 2006;**193**:59–67.

Shapiro B, Rambaut A, Drummond AJ. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 2006;**23**:7–9.

Shi T, McAllister DA, O'Brien KL *et al.* Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *Lancet* 2017;**390**:946–58.

Simoes EA. Respiratory syncytial virus infection. *Lancet* 1999;**354**:847–52.

Smith GJD, Vijaykrishna D, Bahl J *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 2009;**459**:1122–5.

Spann KM, Collins PL, Teng MN. Genetic recombination during coinfection of two mutants of human respiratory syncytial virus. *J Virol* 2003;**77**:11201–11.

Spann KM, Tran K-C, Chi B *et al.* Suppression of the induction of alpha, beta, and lambda interferons by the NS1 and NS2 proteins of human respiratory syncytial virus in human epithelial cells and macrophages [corrected]. *J Virol* 2004, DOI: 10.1128/jvi.78.8.4363-4369.2004.

Steinhauer DA, Domingo E, Holland JJ. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene* 1992;**122**:281–8.

Stern A, Yeh M Te, Zinger T *et al.* The Evolutionary Pathway to Virulence of an RNA Virus. *Cell* 2017;**169**:35-46.e19.

Stokes KL, Chi MH, Sakamoto K *et al.* Differential Pathogenesis of Respiratory Syncytial Virus Clinical Isolates in BALB/c Mice. *J Virol* 2011;**85**:5782–93.

Stokes KL, Currier MG, Sakamoto K *et al.* The Respiratory Syncytial Virus Fusion Protein and Neutrophils Mediate the Airway Mucin Response to Pathogenic Respiratory Syncytial Virus Infection. *J Virol* 2013, DOI: 10.1128/JVI.01347-13.

Suchard MA, Rambaut A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* 2009;**25**:1370–6.

Sullender WM. Respiratory syncytial virus genetic and antigenic diversity. *Clin Microbiol Rev* 2000;**13**:1–15.

Sullender WM, Mufson MA, Anderson LJ *et al.* Genetic Diversity of the Attachment Protein of Subgroup B Respiratory Syncytial Viruses. *J Virol* 1991;**65**:5425–34.

Sullivan BM, Emonet SF, Welch MJ *et al.* Point mutation in the glycoprotein of lymphocytic choriomeningitis virus is necessary for receptor binding, dendritic cell infection, and long-term persistence. *Proc Natl Acad Sci* 2011;**108**:2969–74.

Sunday F. US not Kenya's largest tourism market. *Standard Group Limited*. https://www.standardmedia.co.ke/business/article/2001270358/us-not-kenya-s-largest-tourism-market. Published February 20, 2018.

Sutherland K a, Collins PL, Peeples ME. Synergistic effects of gene-end signal mutations and the M2-1 protein on transcription termination by respiratory syncytial virus. *Virology* 2001;**288**:295–307.

Tamura K, Stecher G, Peterson D *et al.* MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 2013;**30**:2725–9.

Tan L, Coenjaerts FEJ, Houspie L *et al.* The Comparative Genomics of Human Respiratory Syncytial Virus Subgroups A and B: Genetic Variability and Molecular Evolutionary Dynamics. *J Virol* 2013;**87**:8213–26.

Tan L, Lemey P, Houspie L *et al.* Genetic Variability among Complete Human Respiratory Syncytial Virus Subgroup A Genomes: Bridging Molecular Evolutionary Dynamics and Epidemiology. *PLoS One* 2012;**7**:e51439.

Taylor G, Wyld S, Valarcher JF *et al.* Recombinant bovine respiratory syncytial virus with deletion of the SH gene induces increased apoptosis and pro-inflammatory cytokines in vitro, and is attenuated and induces protective immunity in calves. *J Gen Virol* 2014, DOI: 10.1099/vir.0.064931-0.

The Economist. More than minerals. *https://www.economist.com/news/middle-east-and-africa/21574012-chinese-trade-africa-keeps-growing-fears-neocolonialism-are-overdone-more* 2013.

The IMpact-RSV Study Group. Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. *Pediatrics* 1998;**102**:531–7.

The Report: Kenya 2017. *https://www.oxfordbusinessgroup.com/kenya-2017/tourism* 2017.

Toft C, Andersson SGE. Evolutionary microbial genomics: Insights into bacterial host adaptation. *Nat Rev Genet* 2010;**11**:465–75.

Tolley KP, Marriott AC, Simpson A *et al.* Identification of mutations contributing to the reduced virulence of a modified strain of respiratory syncytial virus. *Vaccine* 1996;**14**:1637–46.

Tran DN, Pham TMH, Ha MT *et al.* Molecular Epidemiology and Disease Severity of Human Respiratory Syncytial Virus in Vietnam. Varga SM (ed.). *PLoS One*

2013;**8**:e45436.

Trento A, Casas I, Calderon A *et al.* Ten Years of Global Evolution of the Human Respiratory Syncytial Virus BA Genotype with a 60-Nucleotide Duplication in the G Protein Gene. *J Virol* 2010;**84**:7500–12.

Trento A, Galiano M, Videla C *et al.* Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *J Gen Virol* 2003;**84**:3115–20.

Trento A, Viegas M, Galiano M *et al.* Natural History of Human Respiratory Syncytial Virus Inferred from Phylogenetic Analysis of the Attachment (G) Glycoprotein with a 60-Nucleotide Duplication. *J Virol* 2006;**80**:975–84.

Tsukagoshi H, Yokoi H, Kobayashi M *et al.* Genetic analysis of attachment glycoprotein (G) gene in new genotype ON1 of human respiratory syncytial virus detected in Japan. *Microbiol Immunol* 2013;**57**:655–9.

Valdés O, Martínez I, Valdivia A *et al.* Unusual Antigenic and Genetic Characteristics of Human Respiratory Syncytial Viruses Isolated in Cuba. *J Virol* 1998;**72**:7589–92.

Valley-Omar Z, Muloiwa R, Hu NC *et al.* Novel respiratory syncytial virus subtype ON1 among children, Cape Town, South Africa, 2012. *Emerg Infect Dis* 2013;**19**:668–70.

Venter M, Collinson M, Schoub BD. Molecular epidemiological analysis of community circulating respiratory syncytial virus in rural South Africa: Comparison of viruses and genotypes responsible for different disease manifestations. *J Med Virol* 2002;**68**:452–61.

Venter M, Madhi SA, Tiemessen CT *et al.* Genetic diversity and molecular epidemiology of respiratory syncytial virus over four consecutive seasons in

South Africa: Identification of new subgroup A and B genotypes. *J Gen Virol* 2001;**82**:2117–24.

Viegas M, Goya S, Mistchenko AS. Sixteen years of evolution of human respiratory syncytial virus subgroup A in Buenos Aires, Argentina: GA2 the prevalent genotype through the years. *Infect Genet Evol* 2016;**43**:213–21.

Villenave R, Thavagnanam S, Sarlang S *et al.* In vitro modeling of respiratory syncytial virus infection of pediatric bronchial epithelium, the primary target of infection in vivo. *Proc Natl Acad Sci* 2012, DOI: 10.1073/pnas.1110203109.

Wallace RG, Fitch WM. Influenza A H5N1 immigration is filtered out at some international borders. *PLoS One* 2008;**3**, DOI: 10.1371/journal.pone.0001697.

Wallace RG, HoDac H, Lathrop RH *et al.* A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci* 2007;**104**:4473–8.

Walsh EE, Mcconnochie KM, Long CE *et al.* Severity of Respiratory Syncytial Virus Infection Is Related to Virus Strain. 1991:1989–90.

Walsh EE, McConnochie KM, Long CE *et al.* Severity of respiratory syncytial virus infection is related to virus strain. *J Infect Dis* 1997;**175**:814–20.

Wang TH, Donaldson YK, Brettle RP *et al.* Identification of Shared Populations of Human Immunodeficiency Virus Type 1 Infecting Microglia and Tissue Macrophages outside the Central Nervous System. *J Virol* 2001, DOI: 10.1128/JVI.75.23.11686-11699.2001.

Waris M. Pattern of respiratory syncytial virus epidemics in Finland: Two-year cycles with alternating prevalence of groups A and B. *J Infect Dis* 1991;**163**:464–9.

Webby RJ, Swenson SL, Krauss SL *et al.* Evolution of swine H3N2 influenza viruses in the United States. *J Virol* 2000;**74**:8243–51.

Weber MW, Mulholland EK, Greenwood BM. Respiratory syncytial virus infection in

tropical and developing countries. *Trop Med Int Health* 1998;**3**:268–80.

White LJ, Mandl JN, Gomes MGM *et al.* Understanding the transmission dynamics of respiratory syncytial virus using multiple time series and nested models. *Math Biosci* 2007;**209**:222–39.

White LJ, Waris M, Cane PA *et al.* The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol Infect* 2005;**133**:279–89.

Williams BG, Gouws E, Boschi-Pinto C *et al.* Estimates of world-wide distribution of child deaths from acute respiratory infections. *Lancet Infect Dis* 2002;**2**:25–32.

Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;**15**:R46.

Yamaguchi M, Sano Y, Dapat IC *et al.* High frequency of repeated infections due to emerging genotypes of human respiratory syncytial viruses among children during eight successive epidemic seasons in Japan. *J Clin Microbiol* 2011;**49**:1034–40.

Yang X, Charlebois P, Gnerre S *et al.* De novo assembly of highly diverse viral populations. *BMC Genomics* 2012;**13**:475.

Yoshihara K, Le MN, Okamoto M *et al.* Association of RSV-A ON1 genotype with Increased Pediatric Acute Lower Respiratory Tract Infection in Vietnam. *Sci Rep* 2016;**6**:27856.

Yu X, Kou Y, Xia D *et al.* Human respiratory syncytial virus in children with lower respiratory tract infections or influenza-like illness and its co-infection characteristics with viruses and atypical bacteria in Hangzhou, China. *J Clin Virol* 2015;**69**:1–6.

Zhu Y, Zembower TR, Metzger KE *et al.* Investigation of respiratory syncytial virus

outbreak on an adult stem cell transplant unit by use of whole-genome sequencing. Diekema DJ (ed.). *J Clin Microbiol* 2017;**55**:2956–63.

Zimin A V., Marçais G, Puiu D *et al.* The MaSuRCA genome assembler. *Bioinformatics* 2013;**29**:2669–77.

Zlateva KT, Lemey P, Moës E *et al.* Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein Genetic Variability and Molecular Evolution of the Human Respiratory Syncytial Virus Subgroup B Attachment G Protein. *J Virol* 2005;**79**:9157.

Zlateva KT, Lemey P, Vandamme A *et al.* Molecular Evolution and Circulation Patterns of Human Respiratory Syncytial Virus Subgroup A: Positively Selected Sites in the Attachment G Glycoprotein. *J Virol* 2004;**78**:4675–83.

Zlateva KT, Vijgen L, Dekeersmaeker N *et al.* Subgroup prevalence and genotype circulation patterns of human respiratory syncytial virus in belgium during ten successive epidemic seasons. *J Clin Microbiol* 2007;**45**:3022–30.

7 Appendices

7.1 Study scientific and ethical approval

KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840 - 00200 NAIROBI - Kenya
Tel: (254) (020) 2722541, 254 (020) 2713349, 0722-205901, 0733-400003 Fax (254) (020) 2720030
Email: director@kemri.org info@kemri.org Website:www.kemri.org

KEMRI/RES/7/3/1 **January 27, 2016**

TO: **MR. JAMES RICHARD OTIENO,**
PRINCIPAL INVESTIGATOR

THROUGH: **DR. BENJAMIN TSOFA,**
DIRECTOR, CGMR-C,
KILIFI

DIRECTOR
CENTRE FOR GEOGRAPHIC MEDICINE
RESEARCH, COAST

Dear Sir,

RE: SERU PROTOCOL NO. KEMRI/SERU/CGMR-C/026/3177 (*RESUBMISSION-INITIAL SUBMISSION*): CHARACTERIZING THE GENOMIC DIVERSITY, EVOLUTION AND PHYLOGEOGRAPHY OF RESPIRATORY SYNCYTIAL VIRUS GENOTYPE ON1 IN KENYA

Reference is made to your letter dated 13th January 2016. The KEMRI Scientific and Ethics Review Unit (SERU) acknowledge receipt of the revised study document on 19th January 2016.

This is to inform you that the Committee notes that the issues raised at the 246th B and C joint meeting of the SERU held on 15th December 2015 have been adequately addressed.

Consequently, this study is granted approval for implementation effective this day, **January 27, 2016** to **January 26, 2017**. Please note that authorization to conduct this study will automatically expire on **January 26, 2017.** If you plan to continue data collection or analysis beyond this date, please submit an application for continuation approval to the SERU by **December 15, 2016**.

You are required to submit any proposed changes to this study to the SERU for review and the changes should not be initiated until written approval from the SERU is received. Please note that any unanticipated problems resulting from the implementation of this study should be brought to the attention of the SERU and you should advise the SERU when the study is completed or discontinued.

You may embark on the study.

Yours faithfully,

EAB

PROF. ELIZABETH BUKUSI,
ACTING HEAD,
KEMRI SCIENTIFIC AND ETHICS REVIEW UNIT

RECEIVED
04 FEB 2016
DIRECTOR'S OFFICE

In Search of Better Health

7.2 RSV-A WGS sequencing primers [6-amplicon method]

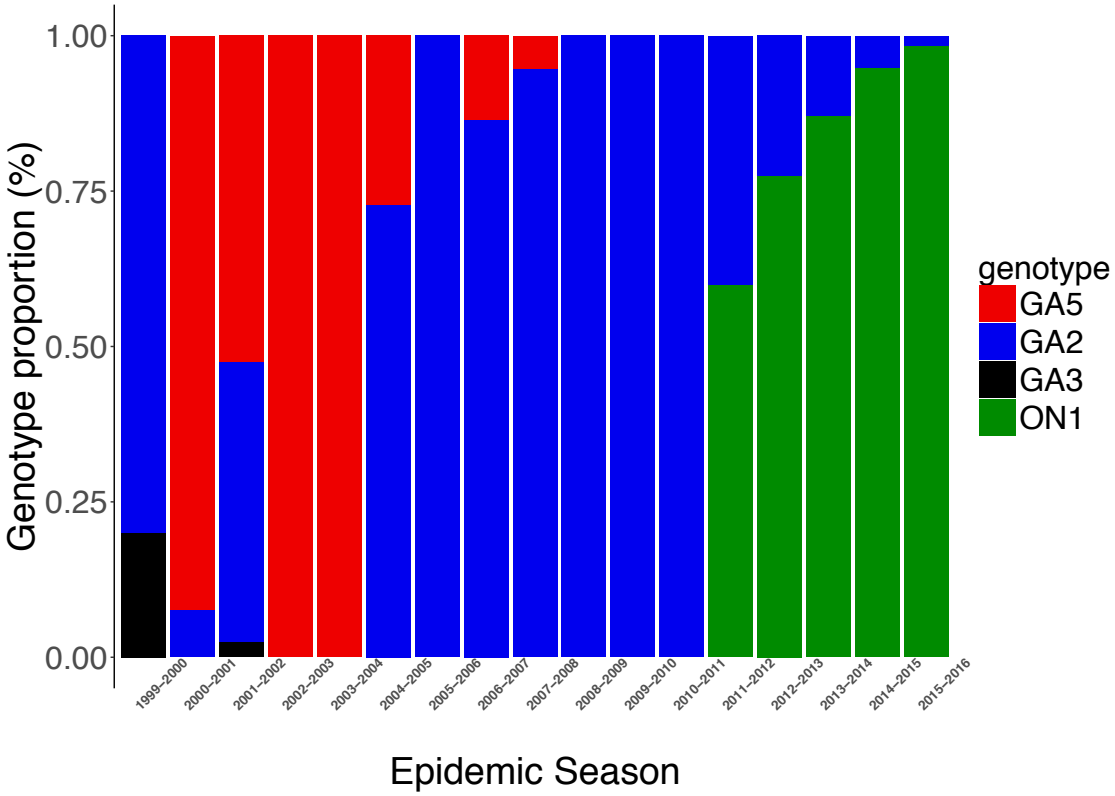
Primer	Amplicon	Sequence	Length	Melting Temperature (Tm)	GC fraction
F_rsvas	1	ACGCGAAAAAATGCGTACAAC	21	57.13	0.43
F_7543	1	TGAATGGCATTGTATTTGTGCATGT	25	58.35	0.36
F_8823	1	TTAACTAATGCTTTGGCTAAGGCAG	25	57.68	0.4
R_70	1	CACGTATGTTTCCATATTTGCCCC	24	58.38	0.46
R_78	1	GGCACCCATATTGTAAGTGATGC	23	57.89	0.48
F_211	2	ATGGCAAAAGACACATCAGATGAAG	25	57.92	0.4
F_751	2	CAGATGAAGTGTCTCTCAATCCAAC	25	57.52	0.44
R_1265	2	CAACTCCATTGTTATTTGCCCCA	23	57.54	0.43
R_1769	2	TGCTAACTGCACTGCATGTTG	21	57.73	0.48
F_2690	3	AGCATATGCAGCAACAATCCAAC	23	58.31	0.43
F_3004	3	TTTGTACCCTGCAGCATATGCA	22	58.42	0.45
R_1939	3	TCAATCAAGTCTTGAGAGGTCCAAT	25	57.68	0.4
R_11745	3	GGTCCAATGGATTTCATTGAATGGT	25	57.84	0.4
F_248	4	AATCAGCATGTGTTGCCATGAG	22	57.74	0.45
F_1734	4	TTTTGAATGGCCACCCCATG	20	56.87	0.5
R_87	4	AGATTGTACACCATGCAGTTCATC	24	57.21	0.42
R_393	4	TGCTAGCAAATAATCTGCTTGAGC	24	57.85	0.42
F_886	5	GGTAGAATGTTTGCAATGCAACC	23	57.28	0.43
F_1850	5	AGTGCTCTATCATCACAGATCTCAG	25	57.55	0.44
R_217	5	AATCTATGTTAACAACCCAAGGGCA	25	58.42	0.4
R_2029	5	CCAAGGGCAAACTGTGAATTCTG	23	58.44	0.48
F_373	6	AGTAGTAGACCATGTGAATTCCCTG	25	57.60	0.44
F_1051	6	GAATTCCCTGCATCAATACCAGC	23	57.88	0.48
R_362	6	CTGAAAACTTCATTACGTCCAGCTA	25	57.23	0.4
R_46462	6	AATACAGTGTTAGTGTGTAGCCATG	25	56.95	0.4
R_rsvae	6	ACGAGAAAAAAAGTGTCAAAAACTAATA	28	55.09	0.25

7.3 RSV-A WGS sequencing primers [14-amplicon method]

Primer	Sequence	Length	Melting Temperature (Tm)	GC fraction	Strand	Position (in plus strand)
A_F_1105	TTGACAATGATGAAGTAGC	19	48.52	0.37	Plus	104
A_F_12259	TTCCATAACAAAACCTTTGA	20	48.63	0.3	Plus	4398
A_F_1378	GGTGTAATAGATACACCTTG	20	48.44	0.4	Plus	6600
A_F_1434	GATGGTACTGTGACAATG	18	48.24	0.44	Plus	6700
A_F_1551	ATATAAGTAAACCAGTCAGAC	21	48.03	0.33	Plus	10963
A_F_2104	GATGTAGAGCTTTGAGTTAA	20	48.28	0.35	Plus	2257
A_F_2361	TATGACATACAAGAGTATGAC	21	47.91	0.33	Plus	8789
A_F_2442	TGAAAGTTTTCTTCAATGC	19	47.78	0.32	Plus	13108
A_F_4189	ATAATCTCCATCATGATTGC	20	48.70	0.35	Plus	4347
A_F_5014	GATGTCAAAGTCTATGCTATA	21	48.25	0.33	Plus	8877
A_F_657	TGGGGAGAGGGATATATAA	19	48.04	0.42	Plus	13068
A_F_7149	TTAACTAATGCTTTGGCTAA	20	48.26	0.3	Plus	163
A_F_774	CATGCTCAAGCAGATTAT	18	47.79	0.39	Plus	11001
A_F_950	GAGAAGATGCAAACAACA	18	48.63	0.39	Plus	2331
A_F_rvas-6	ACGCGAAAAAATGCG	15	49.00	0.47	Minus	15233
A_R_1464	TTGCAGGACCTATTGTAA	18	48.13	0.39	Minus	14559
A_R_2054	ATCATTTCGAGATCCACT	18	48.07	0.39	Minus	10123
A_R_2299	CCCAATCCAATTTTGCTA	18	48.06	0.39	Minus	12408
A_R_2348	ATTCTTCAGTGATGTTTTGA	20	48.52	0.3	Minus	5775
A_R_2372	TCATAGTGAGATCTTTAACTG	21	48.09	0.33	Minus	3613
A_R_238	GACATAGCATATAACATACCT	21	47.93	0.33	Minus	1310
A_R_261	CTTGACGATGTGTTGTTA	18	48.14	0.39	Minus	1402
A_R_346	CTAGGGAAACTAGTCCATAT	20	48.29	0.4	Minus	10180
A_R_350	CCTATATAACTCTCTAGCACT	21	48.67	0.38	Minus	7878
A_R_358	GCAACTCCATTGTTATTTG	19	48.42	0.37	Minus	5691
A_R_5	GGTGTGGTTACATCATATG	19	48.69	0.42	Minus	3536
A_R_663	AATCGATATCATCTTGAGC	19	47.86	0.37	Minus	14460
A_R_74	ACATGCTGATTGTTTAGTTA	20	48.32	0.3	Minus	7912
A_R_833	GCTATAGTGCTTGTTGTATA	20	48.15	0.35	Minus	12229
B_F_143	ATGTGGCATGCTATTAATC	19	48.62	0.37	Plus	1236
B_F_15612	AACATCCGAGTACCTATC	18	48.00	0.44	Plus	5551
B_F_1577	GGTCATTGCTTGAATGG	17	48.72	0.47	Plus	7661
B_F_17	TGCTTCCTTGGCATC	15	48.03	0.53	Plus	13951
B_F_2225	GCCTACTTTAAGGAATGC	18	48.55	0.44	Plus	9956
B_F_249	TACGTGAACAAACTTCAC	18	48.43	0.39	Plus	3220
B_F_551	GAAATAAGTGGAGCTGC	17	48.14	0.47	Plus	7799
B_F_63	ATGGCTCTTAGCAAAGT	17	48.05	0.41	Plus	1096
B_F_638	AATGGTAGATGAAAGACAAG	20	48.44	0.35	Plus	9812
B_F_651	CACTTACAATATGGGTGC	18	48.56	0.44	Plus	3302
B_F_720	ACTGAGATGATGAGGAAAA	19	48.61	0.37	Plus	12012

B_F_84	GATCTTGGTCTTTATCCAATA	21	48.28	0.33	Plus	12142
B_F_85	CACTTTATTGCATGCTTC	18	47.78	0.39	Plus	13939
B_F_8930	AAATCCAGAACACACAAG	18	48.09	0.39	Plus	5398
B_R_1024	GATGGAGGATGTTGCA	16	48.21	0.5	Minus	9121
B_R_1299	TTGCCCTTTATTGATTCTAG	20	48.45	0.35	Minus	2382
B_R_131	AACCAATGTATTAACCATGA	20	47.82	0.3	Minus	9139
B_R_136	TTTGGACATGTTTGCATT	18	48.58	0.33	Minus	4645
B_R_15653	GAATACAGTGTTAGTGTGTA	20	47.96	0.35	Minus	15062
B_R_18507	ATTCCTCCTAGATCAAAATG	20	47.77	0.35	Plus	15281
B_R_2394	GCTTTCTTTGGTTACTTCTA	20	48.74	0.35	Minus	2464
B_R_2432	CTAATTCTTGTGTCAAACTAC	21	48.35	0.33	Minus	11262
B_R_2565	TCTTTATACTAGCTGGGTAAT	21	48.60	0.33	Minus	11172
B_R_260	CCACGATTTTTATTGGATG	19	47.89	0.37	Minus	6970
B_R_359	TGGTAATGTTAAACTGTTCA	20	48.14	0.3	Minus	6809
B_R_541	TGTGTTGGATGATAATCTATG	21	48.73	0.33	Minus	13420
B_R_5492	AACTTGTATAAGCACGATG	19	48.72	0.37	Minus	4734
B_R_778	ACAACCCAAGGGCA	14	48.23	0.57	Minus	13396
B_R_rsvae-9	ACGAGAAAAAAAGTGTCAA	19	48.80	0.32	Minus	15233

7.4 Kilifi RSV-A Genotype patterns 2000-2016



7.5 Genome Details

Sample	Genome Accession	Genotype	PCR strategy	No. of Reads	RSV Reads	Genome length	Mean coverage	Assembler	PCR Ct	Collection date	SRA Accession
KEN/Kilifi/WGS/1022_28/12/2011	MH181878	GA2	6-amplicon	1,270,622	393,958	14,912	4,483.13	VIRALNGS	NA	28/12/11	SAMN08724833
KEN/Kilifi/WGS/1024_09/02/2012	MH181879	GA2	6-amplicon	1,407,764	270,288	14,697	2,961.69	VIRALNGS	22.94	09/02/12	SAMN08724834
KEN/Kilifi/WGS/1025_13/02/2012	MH181908	ON1	14-amplicon	3,293,722	3,088,934	15,168	53,178.19	VIRALNGS	19.26	13/02/12	SAMN08724835
KEN/Kilifi/WGS/1026_16/02/2012	MH181880	GA2	6-amplicon	1,454,390	332,690	14,697	3,457.63	VIRALNGS	21.38	16/02/12	SAMN08724836
KEN/Kilifi/WGS/1028_02/03/2012	MH181881	GA2	6-amplicon	939,372	442,216	14,913	4,740.40	VIRALNGS	19.55	02/03/12	SAMN08724837
KEN/Kilifi/WGS/1029_03/03/2012	MH181882	GA2	6-amplicon	1,125,168	465,202	14,903	5,023.98	VIRALNGS	21.31	03/03/12	SAMN08724838
KEN/Kilifi/WGS/1031_05/03/2012	MH181883	GA2	6-amplicon	1,385,586	1,036,466	14,915	11,068.56	VIRALNGS	20.17	05/03/12	SAMN08724839
KEN/Kilifi/WGS/1032_07/03/2012	MH181909	ON1	14-amplicon	2,602,100	2,252,496	15,070	38,777.92	VIRALNGS	21.64	07/03/12	SAMN08724840
KEN/Kilifi/WGS/1033_09/03/2012	MH181910	ON1	14-amplicon	3,393,018	3,272,224	15,197	55,740.25	VIRALNGS	17.75	09/03/12	SAMN08724841
KEN/Kilifi/WGS/1034_11/03/2012	MH181911	ON1	14-amplicon	2,218,728	1,065,438	14,959	18,236.55	VIRALNGS	24.22	11/03/12	SAMN08724842
KEN/Kilifi/WGS/1035_16/03/2012	MH181912	ON1	14-amplicon	3,380,044	3,262,410	15,180	55,995.24	VIRALNGS	16.98	16/03/12	SAMN08724843
KEN/Kilifi/WGS/1036_21/03/2012	MH181913	ON1	14-amplicon	3,858,536	3,746,792	15,205	64,113.11	VIRALNGS	17.98	21/03/12	SAMN08724844
KEN/Kilifi/WGS/1037_23/03/2012	MH181884	GA2	6-amplicon	984,444	1,626	14,332	44.33	VIRALNGS	24.83	23/03/12	SAMN08724845
KEN/Kilifi/WGS/1038_23/03/2012	MH181914	ON1	14-amplicon	3,886,448	3,641,296	15,152	62,377.42	VIRALNGS	20.41	23/03/12	SAMN08724846
KEN/Kilifi/WGS/1039_25/03/2012	MH181885	GA2	6-amplicon	1,620,128	33,722	14,686	406.46	VIRALNGS	26.31	25/03/12	SAMN08724847
KEN/Kilifi/WGS/1040_26/03/2012	MH181915	ON1	14-amplicon	4,350,992	3,963,646	15,061	66,457.67	VIRALNGS	20.16	26/03/12	SAMN08724848
KEN/Kilifi/WGS/1041_29/03/2012	MH181916	ON1	14-amplicon	3,056,872	2,224,280	15,140	38,552.58	VIRALNGS	21.25	29/03/12	SAMN08724849
KEN/Kilifi/WGS/1043_05/04/2012	MH181917	ON1	14-amplicon	2,191,722	1,876,406	15,232	32,307.51	SPADES	24.16	05/04/12	SAMN08724850
KEN/Kilifi/WGS/1044_10/04/2012	MH181918	ON1	14-amplicon	2,916,484	2,762,130	15,164	47,482.85	VIRALNGS	20.87	10/04/12	SAMN08724851
KEN/Kilifi/WGS/1045_10/04/2012	MH181886	GA2	6-amplicon	1,494,678	958,672	14,891	9,996.56	VIRALNGS	22.16	10/04/12	SAMN08724852
KEN/Kilifi/WGS/1046_13/04/2012	MH181887	GA2	6-amplicon	1,251,788	880,908	14,549	9,824.38	VIRALNGS	21.33	13/04/12	SAMN08724853
KEN/Kilifi/WGS/1047_15/04/2012	MH181888	GA2	6-amplicon	1,641,672	383,808	14,697	4,419.82	VIRALNGS	22.62	15/04/12	SAMN08724854

KEN/Kilifi/WGS/1048_16/04/2012	MH181889	GA2	6-amplicon	1,345,204	269,926	14,669	3,341.61	VIRALNGS	25.2	16/04/12	SAMN08724855
KEN/Kilifi/WGS/1049_17/04/2012	MH181890	GA2	6-amplicon	1,287,742	912,588	14,876	9,912.84	VIRALNGS	20.66	17/04/12	SAMN08724856
KEN/Kilifi/WGS/1050_17/04/2012	MH181891	GA2	6-amplicon	1,483,678	35,648	14,589	431.77	VIRALNGS	25.41	17/04/12	SAMN08724857
KEN/Kilifi/WGS/1051_24/04/2012	MH181919	ON1	14-amplicon	3,036,992	2,853,918	15,198	49,164.32	VIRALNGS	18.44	24/04/12	SAMN08724858
KEN/Kilifi/WGS/1052_25/04/2012	MH181892	GA2	6-amplicon	1,555,690	879,362	14,693	9,480.92	VIRALNGS	19.98	25/04/12	SAMN08724859
KEN/Kilifi/WGS/1053_26/04/2012	MH181920	ON1	14-amplicon	2,552,538	2,084,818	15,116	35,963.84	VIRALNGS	21.81	26/04/12	SAMN08724860
KEN/Kilifi/WGS/1054_16/05/2012	MH181921	ON1	14-amplicon	3,235,132	2,694,000	15,138	45,968.25	VIRALNGS	20.81	16/05/12	SAMN08724861
KEN/Kilifi/WGS/1056_30/05/2012	MH181922	ON1	14-amplicon	2,935,820	1,780,478	15,049	30,814.75	VIRALNGS	22.72	30/05/12	SAMN08724862
KEN/Kilifi/WGS/1058_08/06/2012	MH181923	ON1	14-amplicon	2,312,598	1,906,202	15,106	32,807.75	SPADES	21.04	08/06/12	SAMN08724863
KEN/Kilifi/WGS/1059_08/06/2012	MH181924	ON1	14-amplicon	2,225,810	2,187,994	15,215	37,245.25	VIRALNGS	16.86	08/06/12	SAMN08724864
KEN/Kilifi/WGS/1060_11/06/2012	MH181925	ON1	14-amplicon	2,470,026	2,348,036	15,188	40,483.39	VIRALNGS	18.23	11/06/12	SAMN08724865
KEN/Kilifi/WGS/1061_13/06/2012	MH181926	ON1	14-amplicon	2,071,626	1,772,624	15,049	30,725.88	VIRALNGS	21.53	13/06/12	SAMN08724866
KEN/Kilifi/WGS/1062_15/06/2012	MH181927	ON1	14-amplicon	1,688,810	1,123,748	15,059	19,373.45	VIRALNGS	23.13	15/06/12	SAMN08724867
KEN/Kilifi/WGS/1063_18/06/2012	MH181928	ON1	14-amplicon	2,887,012	2,207,944	15,042	37,782.03	VIRALNGS	21.58	18/06/12	SAMN08724868
KEN/Kilifi/WGS/1064_24/06/2012	MH181893	GA2	6-amplicon	1,151,112	400,912	14,572	4,631.14	VIRALNGS	23.79	24/06/12	SAMN08724869
KEN/Kilifi/WGS/1066_28/06/2012	MH181929	ON1	14-amplicon	2,729,430	2,628,610	15,175	45,411.89	VIRALNGS	18.55	28/06/12	SAMN08724870
KEN/Kilifi/WGS/1067_30/06/2012	MH181894	GA2	6-amplicon	1,691,236	114,412	14,698	1,384.07	VIRALNGS	29.82	30/06/12	SAMN08724871
KEN/Kilifi/WGS/1068_10/07/2012	MH181930	ON1	14-amplicon	2,525,642	2,383,726	15,207	40,719.45	VIRALNGS	20.98	10/07/12	SAMN08724872
KEN/Kilifi/WGS/1070_23/07/2012	MH181931	ON1	14-amplicon	2,159,538	1,535,640	15,049	26,480.83	VIRALNGS	25.32	23/07/12	SAMN08724873
KEN/Kilifi/WGS/1071_14/08/2012	MH181932	ON1	14-amplicon	2,675,604	2,520,028	15,233	43,447.66	SPADES	21.45	14/08/12	SAMN08724874
KEN/Kilifi/WGS/1075_27/10/2012	MH181933	ON1	14-amplicon	2,194,346	1,857,942	15,232	31,956.04	SPADES	24.8	27/10/12	SAMN08724875
KEN/Kilifi/WGS/1076_30/10/2012	MH181934	ON1	14-amplicon	3,041,964	2,314,252	15,068	39,401.72	VIRALNGS	23.24	30/10/12	SAMN08724876
KEN/Kilifi/WGS/1077_31/10/2012	MH181935	ON1	14-amplicon	2,438,426	2,175,418	15,232	37,609.41	SPADES	22.49	31/10/12	SAMN08724877
KEN/Kilifi/WGS/1078_31/10/2012	MH181936	ON1	14-amplicon	2,476,744	2,311,046	15,230	39,698.20	SPADES	21.93	31/10/12	SAMN08724878
KEN/Kilifi/WGS/1079_05/11/2012	MH181937	ON1	14-amplicon	2,114,774	1,819,602	15,232	31,236.65	SPADES	22.13	05/11/12	SAMN08724879
KEN/Kilifi/WGS/1080_05/11/2012	MH181938	ON1	14-amplicon	1,786,686	1,465,770	15,074	25,257.87	VIRALNGS	25.1	05/11/12	SAMN08724880

KEN/Kilifi/WGS/1081_07/11/2012	MH181939	ON1	14-amplicon	2,096,870	1,926,670	15,151	33,319.15	VIRALNGS	21.73	07/11/12	SAMN08724881
KEN/Kilifi/WGS/1082_10/11/2012	MH181940	ON1	14-amplicon	1,699,646	1,240,272	15,214	21,458.29	SPADES	24.24	10/11/12	SAMN08724882
KEN/Kilifi/WGS/1083_12/11/2012	MH181941	ON1	14-amplicon	2,060,860	1,899,520	15,222	32,687.98	SPADES	19.74	12/11/12	SAMN08724883
KEN/Kilifi/WGS/1084_15/11/2012	MH181942	ON1	14-amplicon	1,632,392	380,376	14,959	6,498.76	VIRALNGS	26.16	15/11/12	SAMN08724884
KEN/Kilifi/WGS/1086_15/11/2012	MH181943	ON1	14-amplicon	2,275,828	2,057,472	15,226	35,609.99	SPADES	25.78	15/11/12	SAMN08724885
KEN/Kilifi/WGS/1087_17/11/2012	MH181944	ON1	14-amplicon	2,287,906	2,032,176	15,049	34,971.14	VIRALNGS	21.43	17/11/12	SAMN08724886
KEN/Kilifi/WGS/1088_19/11/2012	MH181945	ON1	14-amplicon	2,701,940	2,645,394	15,171	45,076.37	VIRALNGS	18.92	19/11/12	SAMN08724887
KEN/Kilifi/WGS/1089_19/11/2012	MH181946	ON1	14-amplicon	2,355,836	1,933,388	15,110	33,278.48	VIRALNGS	23.96	19/11/12	SAMN08724888
KEN/Kilifi/WGS/1090_19/11/2012	MH181947	ON1	14-amplicon	2,286,976	2,185,794	15,171	37,746.72	VIRALNGS	18.54	19/11/12	SAMN08724889
KEN/Kilifi/WGS/1091_22/11/2012	MH181948	ON1	14-amplicon	2,359,658	2,288,890	15,198	39,369.58	VIRALNGS	16.91	22/11/12	SAMN08724890
KEN/Kilifi/WGS/1092_23/11/2012	MH181949	ON1	14-amplicon	2,060,978	1,566,076	15,025	26,801.24	VIRALNGS	23.19	23/11/12	SAMN08724891
KEN/Kilifi/WGS/1093_23/11/2012	MH181950	ON1	14-amplicon	1,182,120	416,874	15,150	7,324.31	SPADES	30.01	23/11/12	SAMN08724892
KEN/Kilifi/WGS/1094_23/11/2012	MH181951	ON1	14-amplicon	2,346,184	2,265,498	15,227	39,009.62	SPADES	17.63	23/11/12	SAMN08724893
KEN/Kilifi/WGS/1095_24/11/2012	MH181952	ON1	14-amplicon	2,169,590	2,002,620	15,104	34,587.63	VIRALNGS	24.14	24/11/12	SAMN08724894
KEN/Kilifi/WGS/1096_24/11/2012	MH181953	ON1	14-amplicon	2,697,992	2,627,426	15,183	45,492.81	VIRALNGS	19.92	24/11/12	SAMN08724895
KEN/Kilifi/WGS/1097_24/11/2012	MH181954	ON1	14-amplicon	2,698,368	2,339,798	15,231	40,334.68	SPADES	21.98	24/11/12	SAMN08724896
KEN/Kilifi/WGS/1098_25/11/2012	MH181955	ON1	14-amplicon	2,671,940	2,502,752	15,232	43,168.08	SPADES	23.48	25/11/12	SAMN08724897
KEN/Kilifi/WGS/1099_25/11/2012	MH181956	ON1	14-amplicon	2,654,944	2,542,430	15,183	43,877.84	VIRALNGS	20.64	25/11/12	SAMN08724898
KEN/Kilifi/WGS/1100_25/11/2012	MH181957	ON1	14-amplicon	2,232,292	2,011,214	15,033	34,492.22	VIRALNGS	21.98	25/11/12	SAMN08724899
KEN/Kilifi/WGS/1101_25/11/2012	MH181895	GA2	6-amplicon	1,476,584	531,814	14,883	6,072.07	VIRALNGS	24.75	25/11/12	SAMN08724900
KEN/Kilifi/WGS/1102_27/11/2012	MH181958	ON1	14-amplicon	3,137,122	3,001,672	15,185	51,523.50	VIRALNGS	20.18	27/11/12	SAMN08724901
KEN/Kilifi/WGS/1103_27/11/2012	MH181959	ON1	14-amplicon	2,561,480	2,387,270	15,145	41,078.44	VIRALNGS	20.05	27/11/12	SAMN08724902
KEN/Kilifi/WGS/1104_27/11/2012	MH181960	ON1	14-amplicon	2,367,006	2,217,710	15,139	38,181.27	VIRALNGS	22.13	27/11/12	SAMN08724903
KEN/Kilifi/WGS/1105_28/11/2012	MH181961	ON1	14-amplicon	2,553,394	2,370,212	15,192	40,462.29	SPADES	21.72	28/11/12	SAMN08724904
KEN/Kilifi/WGS/1106_28/11/2012	MH181962	ON1	14-amplicon	2,958,126	2,940,148	15,176	50,168.19	VIRALNGS	18.98	28/11/12	SAMN08724905
KEN/Kilifi/WGS/1107_28/11/2012	MH181963	ON1	14-amplicon	2,781,054	1,786,998	15,030	31,236.67	VIRALNGS	27.6	28/11/12	SAMN08724906

KEN/Kilifi/WGS/1108_28/11/2012	MH181896	GA2	6-amplicon	1,553,924	611,370	14,949	7,511.50	VIRALNGS	25.74	28/11/12	SAMN08724907
KEN/Kilifi/WGS/1109_29/11/2012	MH181964	ON1	14-amplicon	3,244,876	3,065,026	15,232	52,644.38	SPADES	27.01	29/11/12	SAMN08724908
KEN/Kilifi/WGS/1110_29/11/2012	MH181965	ON1	14-amplicon	2,750,196	2,413,804	15,147	41,452.34	VIRALNGS	22.03	29/11/12	SAMN08724909
KEN/Kilifi/WGS/1112_29/11/2012	MH181966	ON1	14-amplicon	3,284,922	2,341,218	15,049	40,716.26	VIRALNGS	28.78	29/11/12	SAMN08724910
KEN/Kilifi/WGS/1113_03/12/2012	MH181967	ON1	14-amplicon	3,211,012	3,101,106	15,228	53,002.60	SPADES	21.08	03/12/12	SAMN08724911
KEN/Kilifi/WGS/1114_03/12/2012	MH181968	ON1	14-amplicon	2,226,714	1,582,202	15,050	28,471.78	VIRALNGS	27.02	03/12/12	SAMN08724912
KEN/Kilifi/WGS/1115_04/12/2012	MH181969	ON1	14-amplicon	2,465,632	1,805,348	15,082	31,377.33	VIRALNGS	25	04/12/12	SAMN08724913
KEN/Kilifi/WGS/1117_10/12/2012	MH181970	ON1	14-amplicon	2,254,784	1,873,994	15,232	32,492.81	SPADES	22.1	10/12/12	SAMN08724914
KEN/Kilifi/WGS/1118_12/12/2012	MH181971	ON1	14-amplicon	2,783,040	2,095,178	15,049	36,339.50	VIRALNGS	22.79	12/12/12	SAMN08724915
KEN/Kilifi/WGS/1119_15/12/2012	MH181972	ON1	14-amplicon	2,354,872	2,156,166	15,152	37,309.87	VIRALNGS	20.03	15/12/12	SAMN08724916
KEN/Kilifi/WGS/1120_19/12/2012	MH181973	ON1	14-amplicon	2,108,962	1,920,920	15,099	32,979.37	VIRALNGS	19.74	19/12/12	SAMN08724917
KEN/Kilifi/WGS/1121_21/12/2012	MH181974	ON1	14-amplicon	3,135,066	2,966,696	15,170	51,695.42	VIRALNGS	20.48	21/12/12	SAMN08724918
KEN/Kilifi/WGS/1122_23/12/2012	MH181975	ON1	14-amplicon	2,404,842	2,248,908	15,145	38,581.46	VIRALNGS	18.57	23/12/12	SAMN08724919
KEN/Kilifi/WGS/1124_24/12/2012	MH181976	ON1	14-amplicon	1,571,860	798,782	15,049	13,968.73	VIRALNGS	23.75	24/12/12	SAMN08724920
KEN/Kilifi/WGS/1125_26/12/2012	MH181977	ON1	14-amplicon	2,626,156	2,516,622	15,227	43,345.79	SPADES	16.88	26/12/12	SAMN08724921
KEN/Kilifi/WGS/1126_03/01/2013	MH181978	ON1	14-amplicon	1,969,924	1,436,248	14,959	24,636.93	VIRALNGS	23.18	03/01/13	SAMN08724922
KEN/Kilifi/WGS/1128_07/01/2013	MH181979	ON1	14-amplicon	1,733,862	1,428,270	15,075	24,776.24	VIRALNGS	23.35	07/01/13	SAMN08724923
KEN/Kilifi/WGS/1129_08/01/2013	MH181980	ON1	14-amplicon	2,637,728	2,557,822	15,169	43,864.26	VIRALNGS	20.43	08/01/13	SAMN08724924
KEN/Kilifi/WGS/1131_10/01/2013	MH181981	ON1	14-amplicon	2,778,546	2,598,338	15,114	44,738.50	VIRALNGS	26.33	10/01/13	SAMN08724925
KEN/Kilifi/WGS/1132_11/01/2013	MH181982	ON1	14-amplicon	2,898,380	2,801,846	15,322	48,134.24	SPADES	19.8	11/01/13	SAMN08724926
KEN/Kilifi/WGS/1133_11/01/2013	MH181983	ON1	14-amplicon	3,266,658	3,046,788	15,150	52,451.71	VIRALNGS	22.94	11/01/13	SAMN08724927
KEN/Kilifi/WGS/1134_12/01/2013	MH181984	ON1	14-amplicon	2,704,858	2,439,372	15,118	41,877.57	VIRALNGS	22.01	12/01/13	SAMN08724928
KEN/Kilifi/WGS/1138_14/01/2013	MH181897	GA2	6-amplicon	1,980,492	1,268,738	14,985	15,156.29	VIRALNGS	20.06	14/01/13	SAMN08724929
KEN/Kilifi/WGS/1140_17/01/2013	MH181985	ON1	14-amplicon	2,211,052	1,873,618	15,123	32,450.84	VIRALNGS	21.47	17/01/13	SAMN08724930
KEN/Kilifi/WGS/1141_17/01/2013	MH181986	ON1	14-amplicon	2,438,632	2,291,448	15,160	39,710.23	VIRALNGS	22.23	17/01/13	SAMN08724931
KEN/Kilifi/WGS/1142_21/01/2013	MH181987	ON1	14-amplicon	2,399,228	2,032,498	15,136	35,144.81	VIRALNGS	23.53	21/01/13	SAMN08724932

KEN/Kilifi/WGS/1144_24/01/2013	MH181988	ON1	6-amplicon	996,732	167,134	14,906	2,385.06	VIRALNGS	20.97	24/01/13	SAMN08724933
KEN/Kilifi/WGS/1146_03/02/2013	MH181898	GA2	6-amplicon	1,460,748	247,094	14,693	2,739.55	VIRALNGS	19.26	03/02/13	SAMN08724934
KEN/Kilifi/WGS/1148_07/02/2013	MH181899	GA2	6-amplicon	1,597,584	1,143,202	14,984	1,414.55	VIRALNGS	20.56	07/02/13	SAMN08724935
KEN/Kilifi/WGS/1149_07/02/2013	MH181900	GA2	6-amplicon	1,333,010	777,046	14,978	8,232.99	VIRALNGS	20.88	07/02/13	SAMN08724936
KEN/Kilifi/WGS/1151_10/02/2013	MH181989	ON1	6-amplicon	1,727,108	159,996	15,037	1,869.62	VIRALNGS	18.67	10/02/13	SAMN08724937
KEN/Kilifi/WGS/1153_12/03/2013	MH181990	ON1	6-amplicon	1,520,970	887,602	15,001	10,661.08	VIRALNGS	18.79	12/03/13	SAMN08724938
KEN/Kilifi/WGS/1155_13/03/2013	MH181901	GA2	6-amplicon	1,024,466	779,830	14,985	10,443.44	VIRALNGS	24.58	24/05/13	SAMN08724939
KEN/Kilifi/WGS/1159_24/05/2013	MH181902	GA2	6-amplicon	1,795,530	959,456	14,981	11,739.93	VIRALNGS	22.76	24/05/13	SAMN08724940
KEN/Kilifi/WGS/1160_31/05/2013	MH181903	GA2	6-amplicon	1,349,884	520,174	14,764	6,539.51	VIRALNGS	22.57	31/05/13	SAMN08724941
KEN/Kilifi/WGS/1162_12/10/2013	MH181991	ON1	6-amplicon	1,003,770	784,072	15,047	10,930.69	VIRALNGS	21.5	12/10/13	SAMN08724942
KEN/Kilifi/WGS/1163_15/10/2013	MH181992	ON1	6-amplicon	1,071,636	612,044	15,053	8,500.36	VIRALNGS	22.21	15/10/13	SAMN08724943
KEN/Kilifi/WGS/1166_28/10/2013	MH181993	ON1	6-amplicon	845,020	193,328	15,021	2,668.08	VIRALNGS	26.99	28/10/13	SAMN08724944
KEN/Kilifi/WGS/1167_04/11/2013	MH181994	ON1	6-amplicon	1,165,720	474,594	15,023	6,649.96	VIRALNGS	25.14	04/11/13	SAMN08724945
KEN/Kilifi/WGS/1168_07/11/2013	MH181995	ON1	6-amplicon	1,209,356	203,814	14,979	2,880.96	VIRALNGS	23.27	07/11/13	SAMN08724946
KEN/Kilifi/WGS/1169_07/11/2013	MH181904	GA2	6-amplicon	1,606,550	1,029,916	15,011	10,948.78	VIRALNGS	19.27	07/11/13	SAMN08724947
KEN/Kilifi/WGS/1171_12/11/2013	MH181996	ON1	6-amplicon	947,458	479,690	14,828	6,448.30	VIRALNGS	19.7	12/11/13	SAMN08724948
KEN/Kilifi/WGS/1173_19/11/2013	MH181997	ON1	6-amplicon	1,058,828	662,216	15,052	9,348.85	VIRALNGS	30.66	19/11/13	SAMN08724949
KEN/Kilifi/WGS/1582_01/04/2014	MH181998	ON1	6-amplicon	1,568,332	687,440	15,026	8,437.15	VIRALNGS	24.16	13/04/14	SAMN08724950
KEN/Kilifi/WGS/1583_13/04/2014	MH181999	ON1	6-amplicon	302,368	39,148	14,734	510.27	VIRALNGS	26.44	15/04/14	SAMN08724951
KEN/Kilifi/WGS/1215_15/04/2014	MH181905	GA2	6-amplicon	1,729,644	998,364	15,044	11,774.25	VIRALNGS	21.29	15/04/14	SAMN08724952
KEN/Kilifi/WGS/1584_23/04/2014	MH182000	ON1	6-amplicon	1,842,612	244,414	14,868	2,917.30	VIRALNGS	26.07	26/11/14	SAMN08724953
KEN/Kilifi/WGS/1246_26/11/2014	MH182001	ON1	6-amplicon	1,174,582	955,666	15,187	12,678.17	VIRALNGS	21.34	26/11/14	SAMN08724954
KEN/Kilifi/WGS/1247_04/12/2014	MH182002	ON1	6-amplicon	706,526	491,376	15,055	6,409.80	VIRALNGS	24.28	04/12/14	SAMN08724955
KEN/Kilifi/WGS/1249_05/12/2014	MH182003	ON1	6-amplicon	2,218,820	1,646,984	15,057	19,674.11	VIRALNGS	24.08	05/12/14	SAMN08724956
KEN/Kilifi/WGS/1250_06/12/2014	MH182004	ON1	6-amplicon	218,950	110,400	14,825	1,493.04	VIRALNGS	24.93	06/12/14	SAMN08724957
KEN/Kilifi/WGS/1251_06/12/2014	MH182005	ON1	6-amplicon	1,174,208	89,898	14,802	1,320.67	VIRALNGS	26.32	06/12/14	SAMN08724958

KEN/Kilifi/WGS/1252_07/12/2014	MH182006	ON1	6-amplicon	1,114,664	424,006	15,043	5,840.78	VIRALNGS	24.72	07/12/14	SAMN08724959
KEN/Kilifi/WGS/1253_08/12/2014	MH182007	ON1	6-amplicon	1,657,878	791,536	15,047	10,202.00	VIRALNGS	27.31	08/12/14	SAMN08724960
KEN/Kilifi/WGS/1254_08/12/2014	MH182008	ON1	6-amplicon	2,111,452	1,448,258	15,052	19,567.76	VIRALNGS	24.71	08/12/14	SAMN08724961
KEN/Kilifi/WGS/1256_09/12/2014	MH182009	ON1	6-amplicon	929,072	446,432	15,167	5,904.01	VIRALNGS	27.5	09/12/14	SAMN08724962
KEN/Kilifi/WGS/1257_10/12/2014	MH182010	ON1	6-amplicon	905,314	679,084	15,023	9,435.24	VIRALNGS	24.48	10/12/14	SAMN08724963
KEN/Kilifi/WGS/1258_11/12/2014	MH182011	ON1	6-amplicon	1,549,978	1,255,922	15,058	16,074.53	VIRALNGS	21.74	11/12/14	SAMN08724964
KEN/Kilifi/WGS/1260_14/12/2014	MH182012	ON1	6-amplicon	1,050,160	881,444	15,060	11,114.68	VIRALNGS	23.38	14/12/14	SAMN08724965
KEN/Kilifi/WGS/1261_14/12/2014	MH182013	ON1	6-amplicon	1,109,236	931,816	15,047	12,570.43	VIRALNGS	25.45	14/12/14	SAMN08724966
KEN/Kilifi/WGS/1263_15/12/2014	MH182014	ON1	6-amplicon	1,227,836	608,602	15,180	8,407.71	VIRALNGS	26.53	15/12/14	SAMN08724967
KEN/Kilifi/WGS/1264_15/12/2014	MH182015	ON1	6-amplicon	713,002	11,594	14,172	179.84	VIRALNGS	32.25	15/12/14	SAMN08724968
KEN/Kilifi/WGS/1266_17/12/2014	MH182016	ON1	6-amplicon	1,620,164	1,332,364	15,059	16,697.38	VIRALNGS	23.34	17/12/14	SAMN08724969
KEN/Kilifi/WGS/1268_17/12/2014	MH182017	ON1	6-amplicon	2,064,268	1,740,450	15,148	21,533.23	VIRALNGS	22.85	17/12/14	SAMN08724970
KEN/Kilifi/WGS/1269_17/12/2014	MH182018	ON1	6-amplicon	1,193,216	830,040	15,187	11,270.51	VIRALNGS	25.02	17/12/14	SAMN08724971
KEN/Kilifi/WGS/1271_18/12/2014	MH182019	ON1	6-amplicon	966,686	822,958	15,098	11,699.58	VIRALNGS	22.74	18/12/14	SAMN08724972
KEN/Kilifi/WGS/1273_19/12/2014	MH182020	ON1	6-amplicon	1,279,716	942,368	15,053	12,026.06	VIRALNGS	24.03	19/12/14	SAMN08724973
KEN/Kilifi/WGS/1274_21/12/2014	MH182021	ON1	6-amplicon	1,097,012	811,778	15,184	10,575.22	VIRALNGS	26.9	21/12/14	SAMN08724974
KEN/Kilifi/WGS/1275_22/12/2014	MH182022	ON1	6-amplicon	1,650,292	1,292,028	15,047	17,744.98	VIRALNGS	25.17	22/12/14	SAMN08724975
KEN/Kilifi/WGS/1278_25/12/2014	MH182023	ON1	6-amplicon	1,064,228	735,742	15,155	9,942.56	VIRALNGS	28.76	25/12/14	SAMN08724976
KEN/Kilifi/WGS/1279_27/12/2014	MH182024	ON1	6-amplicon	757,922	130,334	14,824	1,797.32	VIRALNGS	27.91	27/12/14	SAMN08724977
KEN/Kilifi/WGS/1281_30/12/2014	MH182025	ON1	6-amplicon	921,080	610,496	15,164	8,271.09	VIRALNGS	22.41	30/12/14	SAMN08724978
KEN/Kilifi/WGS/1282_31/12/2014	MH182026	ON1	6-amplicon	1,648,622	676,458	15,020	8,667.22	VIRALNGS	29.32	31/12/14	SAMN08724979
KEN/Kilifi/WGS/1283_31/12/2014	MH182027	ON1	6-amplicon	1,315,628	694,854	15,031	9,598.81	VIRALNGS	24.97	31/12/14	SAMN08724980
KEN/Kilifi/WGS/1284_02/01/2015	MH182028	ON1	6-amplicon	1,059,850	813,254	15,184	10,727.02	VIRALNGS	24.28	02/01/15	SAMN08724981
KEN/Kilifi/WGS/1286_04/01/2015	MH182029	ON1	6-amplicon	2,048,806	1,502,422	15,056	18,679.43	VIRALNGS	25.14	04/01/15	SAMN08724982
KEN/Kilifi/WGS/1287_06/01/2015	MH182030	ON1	6-amplicon	1,320,088	1,006,680	15,204	13,479.43	VIRALNGS	22.91	06/01/15	SAMN08724983
KEN/Kilifi/WGS/1288_07/01/2015	MH182031	ON1	6-amplicon	537,376	426,692	15,059	5,429.30	VIRALNGS	23.91	07/01/15	SAMN08724984

KEN/Kilifi/WGS/1289_08/01/2015	MH182032	ON1	6-amplicon	1,099,200	783,598	15,044	11,047.58	VIRALNGS	23.81	08/01/15	SAMN08724985
KEN/Kilifi/WGS/1290_08/01/2015	MH182033	ON1	6-amplicon	930,504	107,914	14,796	1,529.49	VIRALNGS	34.87	08/01/15	SAMN08724986
KEN/Kilifi/WGS/1292_09/01/2015	MH182034	ON1	6-amplicon	770,526	7,412	14,216	111.73	VIRALNGS	28.13	09/01/15	SAMN08724987
KEN/Kilifi/WGS/1293_24/01/2015	MH182035	ON1	6-amplicon	1,040,044	848,258	15,056	10,945.92	VIRALNGS	24.24	24/01/15	SAMN08724988
KEN/Kilifi/WGS/1295_27/01/2015	MH182036	ON1	6-amplicon	959,732	725,710	15,206	9,327.55	VIRALNGS	23.86	27/01/15	SAMN08724989
KEN/Kilifi/WGS/1296_29/01/2015	MH182037	ON1	6-amplicon	1,024,190	681,324	15,052	9,406.83	VIRALNGS	25.01	29/01/15	SAMN08724990
KEN/Kilifi/WGS/1298_31/01/2015	MH182038	ON1	6-amplicon	765,634	523,320	15,047	7,236.79	VIRALNGS	24.66	31/01/15	SAMN08724991
KEN/Kilifi/WGS/1299_04/02/2015	MH182039	ON1	6-amplicon	1,390,092	34,784	14,741	497.77	VIRALNGS	31.57	04/02/15	SAMN08724992
KEN/Kilifi/WGS/1301_08/02/2015	MH182040	ON1	6-amplicon	1,184,664	633,118	15,117	8,281.10	VIRALNGS	25.06	08/02/15	SAMN08724993
KEN/Kilifi/WGS/1302_13/02/2015	MH182041	ON1	6-amplicon	1,134,242	469,232	14,848	6,827.43	VIRALNGS	25.69	13/02/15	SAMN08724994
KEN/Kilifi/WGS/1305_21/02/2015	MH182042	ON1	6-amplicon	912,242	614,448	15,102	8,643.56	VIRALNGS	28.16	21/02/15	SAMN08724995
KEN/Kilifi/WGS/1307_26/02/2015	MH182043	ON1	6-amplicon	1,032,236	867,040	15,033	11,866.93	VIRALNGS	24.37	26/02/15	SAMN08724996
KEN/Kilifi/WGS/1308_05/03/2015	MH182044	ON1	6-amplicon	1,204,456	1,053,810	15,201	13,377.58	VIRALNGS	24.26	05/03/15	SAMN08724997
KEN/Kilifi/WGS/1309_08/03/2015	MH182045	ON1	6-amplicon	1,110,724	915,046	15,053	12,126.92	VIRALNGS	24.14	08/03/15	SAMN08724998
KEN/Kilifi/WGS/1311_21/03/2015	MH182046	ON1	6-amplicon	307,420	193,146	14,740	2,480.30	VIRALNGS	27.9	21/03/15	SAMN08724999
KEN/Kilifi/WGS/1312_22/03/2015	MH182047	ON1	6-amplicon	1,722,762	1,296,644	15,050	15,953.55	VIRALNGS	24.91	22/03/15	SAMN08725000
KEN/Kilifi/WGS/1313_27/03/2015	MH182048	ON1	6-amplicon	1,143,280	429,554	14,827	5,977.84	VIRALNGS	27.51	27/03/15	SAMN08725001
KEN/Kilifi/WGS/1314_31/03/2015	MH181906	GA2	6-amplicon	1,625,342	1,010,186	15,022	11,311.20	VIRALNGS	20.75	31/03/15	SAMN08725002
KEN/Kilifi/WGS/1315_01/04/2015	MH182049	ON1	6-amplicon	1,107,120	740,496	15,066	9,340.28	VIRALNGS	23.95	01/04/15	SAMN08725003
KEN/Kilifi/WGS/1318_08/04/2015	MH182050	ON1	6-amplicon	900,852	759,034	15,202	9,913.31	VIRALNGS	23.52	08/04/15	SAMN08725004
KEN/Kilifi/WGS/1320_12/04/2015	MH182051	ON1	6-amplicon	1,698,698	1,185,294	15,048	14,503.36	VIRALNGS	24.15	12/04/15	SAMN08725005
KEN/Kilifi/WGS/1321_17/04/2015	MH182052	ON1	6-amplicon	1,074,506	944,630	15,019	12,360.20	VIRALNGS	25.73	17/04/15	SAMN08725006
KEN/Kilifi/WGS/1322_17/04/2014	MH181907	GA2	6-amplicon	1,398,136	367,740	14,757	4,605.53	VIRALNGS	23.73	17/04/14	SAMN08725007
KEN/Kilifi/WGS/1297_29/01/2015	MH182053	ON1	6-amplicon	1,598,292	1,046	13,966	39.95	VIRALNGS	NA	29/01/15	SAMN08725008
KEN/Kilifi/WGS/1585_17/02/2016	MH182054	ON1	6-amplicon	1,149,120	700,486	15,082	8,828.96	VIRALNGS	22.09	17/02/16	SAMN08725009
KEN/Kilifi/WGS/1586_28/01/2016	MH182055	ON1	6-amplicon	1,156,030	266,436	14,704	3,938.20	VIRALNGS	22.58	28/01/16	SAMN08725010

KEN/Kilifi/WGS/1587_08/02/2016	MH182056	ON1	6-amplicon	939,876	48,486	14,882	712.62	VIRALNGS	25.82	08/02/16	SAMN08725011
KEN/Kilifi/WGS/1588_29/03/2016	MH182057	ON1	6-amplicon	1,021,102	391,966	14,993	4,937.68	VIRALNGS	25.32	29/03/16	SAMN08725012
KEN/Kilifi/WGS/1589_26/01/2016	MH182058	ON1	6-amplicon	1,281,062	99,496	14,660	1,353.70	VIRALNGS	25.92	26/01/16	SAMN08725013
KEN/Kilifi/WGS/1590_01/02/2016	MH182059	ON1	6-amplicon	1,153,340	441,172	15,035	5,304.08	VIRALNGS	22.02	01/02/16	SAMN08725014
KEN/Kilifi/WGS/1591_18/01/2016	MH182060	ON1	6-amplicon	1,124,140	191,242	15,026	2,565.53	VIRALNGS	22.5	18/01/16	SAMN08725015
KEN/Kilifi/WGS/1592_07/04/2016	MH182061	ON1	6-amplicon	1,089,038	647,740	15,078	8,421.34	VIRALNGS	27.7	07/04/16	SAMN08725016

7.6 SNPs identified from dataset of all Kilifi genomes

Highlighted in yellow are the signature substitutions that differentiate genotype ON1 viruses from GA2 viruses

Genome Position	CDS	Nt Pos.	AA Pos.	Codon Pos.	Nt Change	Variant Frequency	Coverage	Polymorphism	AA Change
2					-AATTTGA	1.0% -> 1.3%	75 -> 96	Deletion	
9					T -> C/-	7.3%/1.0%	96	Mixture	
10					-AA	1.00%	97 -> 99	Deletion	
80	NS1	37	13	1	T -> C	1.50%	137	transition	
112	NS1	69	23	3	A -> G	2.60%	151	transition	
134	NS1	91	31	1	A -> G	2.70%	149	transition	T -> A
172	NS1	129	43	3	G -> A	6.30%	144	transition	
175	NS1	132	44	3	A -> T	3.50%	143	transversion	
187	NS1	144	48	3	A -> G	1.40%	142	transition	
196	NS1	153	51	3	G -> A	1.40%	139	transition	
244	NS1	201	67	3	T -> C	1.70%	172	transition	
359	NS1	316	106	1	G -> T	1.10%	178	transversion	D -> Y
460	NS1	417	139	3	A -> G	1.70%	177	transition	
498					C -> A	2.20%	178	transversion	
507					T -> C	2.80%	179	transition	
518					+AAA	1.70%	181	Insertion	
531					A -> G	3.40%	177	transition	
532					G -> A	4.00%	177	transition	
534					T -> C	5.10%	177	transition	
561					A -> G	99.40%	177	transition	
567					C -> T	1.70%	176	transition	
571					C -> T	1.10%	177	transition	
605	NS2	33	11	3	A -> G	2.30%	176	transition	
629	NS2	57	19	3	A -> G	1.70%	176	transition	
638	NS2	66	22	3	A -> G	82.40%	176	transition	
641	NS2	69	23	3	T -> G	5.10%	176	transversion	
647	NS2	75	25	3	T -> C	2.30%	177	transition	
660	NS2	88	30	1	T -> C	88.80%	178	transition	
734	NS2	162	54	3	A -> G	1.10%	179	transition	
752	NS2	180	60	3	A -> G	2.20%	179	transition	
761	NS2	189	63	3	C -> T	1.10%	179	transition	
770	NS2	198	66	3	A -> G	1.10%	180	transition	
780	NS2	208	70	1	T -> C	1.10%	180	transition	
788	NS2	216	72	3	A -> G	1.10%	180	transition	
800	NS2	228	76	3	T -> C	1.10%	180	transition	
863	NS2	291	97	3	C -> T	3.30%	181	transition	
879	NS2	307	103	1	T -> C	6.60%	182	transition	
896	NS2	324	108	3	T -> C	1.10%	182	transition	
902	NS2	330	110	3	T -> C	1.10%	182	transition	
956					A -> C/T	47.0%/36.5 %	181	SNP	
978					T -> C	5.60%	180	transition	
981					C -> T	1.10%	180	transition	
983					A -> G	81.10%	180	transition	
985					T -> C	1.10%	180	transition	
986					T -> C	1.70%	180	transition	
993					T -> C	1.10%	180	transition	
1008					G -> A	1.10%	182	transition	
1012					A -> G	1.10%	182	transition	
1017					G -> A	4.40%	182	transition	
1024					A -> G	1.10%	181	transition	
1064					A -> G	1.60%	183	transition	
1189	N	105	35	3	T -> A	36.50%	181	transversion	
1204	N	120	40	3	G -> T	6.60%	182	transversion	
1231	N	147	49	3	C -> T	36.50%	181	transition	
1235	N	151	51	1	T -> C	28.20%	181	transition	
1255	N	171	57	3	T -> C	2.80%	181	transition	
1267	N	183	61	3	C -> T	37.40%	182	transition	
1286	N	202	68	1	T -> C	4.40%	182	transition	
1300	N	216	72	3	T -> C	1.10%	182	transition	
1342	N	258	86	3	G -> A	1.10%	181	transition	

1495	N	411	137	3	A -> G	1.10%	182	transition	
1501	N	417	139	3	A -> G	1.10%	182	transition	
1522	N	438	146	3	T -> C	1.10%	182	transition	
1543	N	459	153	3	T -> C	9.90%	182	transition	
1615	N	531	177	3	T -> A	88.50%	183	transversion	
1630	N	546	182	3	G -> A	1.10%	183	transition	
1642	N	558	186	3	T -> C	3.30%	183	transition	
1645	N	561	187	3	T -> C	2.20%	183	transition	
1675	N	591	197	3	T -> C	1.10%	183	transition	
1699	N	615	205	3	A -> C	1.60%	183	transversion	
1720	N	636	212	3	G -> A	1.10%	183	transition	
1765	N	681	227	3	T -> C	3.30%	183	transition	
1801	N	717	239	3	T -> C	2.20%	183	transition	
1807	N	723	241	3	G -> A	6.60%	183	transition	
1813	N	729	243	3	T -> C	1.10%	183	transition	
1820	N	736	246	1	T -> C	2.20%	183	transition	
1825	N	741	247	3	T -> C	1.10%	183	transition	
1827	N	743	248	2	T -> A	3.30%	183	transversion	M -> K
1859	N	775	259	1	C -> A	2.70%	184	transversion	
1861	N	777	259	3	G -> A	1.60%	184	transition	
1876	N	792	264	3	A -> G	1.60%	184	transition	
1885	N	801	267	3	T -> A	1.10%	184	transversion	
1885	N	801	267	3	T -> C	1.10%	184	transition	
1888	N	804	268	3	A -> G	4.90%	184	transition	
1921	N	837	279	3	A -> T	6.50%	184	transversion	
2032	N	948	316	3	A -> G	1.60%	184	transition	
2062	N	978	326	3	C -> T	1.10%	184	transition	
2080	N	996	332	3	C -> T	1.10%	184	transition	
2146	N	1062	354	3	T -> C	1.10%	184	transition	
2170	N	1086	362	3	G -> A	2.20%	184	transition	
2191	N	1107	369	3	C -> T	83.70%	184	transition	
2200	N	1116	372	3	A -> G	3.80%	184	transition	
2267					(A)6 -> (A)7	4.90%	184	Insertion (tandem repeat)	
2267					(A)6 -> (A)5	2.70%	183	Deletion (tandem repeat)	
2272					A -> G	77.00%	183	transition	
2273					G -> A	8.20%	183	transition	
2366	P	75	25	3	G -> A	1.10%	182	transition	
2478	P	187	63	1	A -> G	37.70%	183	transition	I -> V
2495	P	204	68	3	G -> A	5.40%	184	transition	
2513	P	222	74	3	G -> A	1.10%	184	transition	
2522	P	231	77	3	T -> C	98.90%	184	transition	
2525	P	234	78	3	T -> C	3.30%	184	transition	
2585	P	294	98	3	C -> T	98.40%	184	transition	
2588	P	297	99	3	A -> T	5.40%	184	transversion	
2600	P	309	103	3	A -> G	2.70%	184	transition	
2663	P	372	124	3	T -> C	1.10%	184	transition	
2678	P	387	129	3	T -> C	1.60%	184	transition	
2753	P	462	154	3	A -> G	1.10%	184	transition	
2756	P	465	155	3	G -> A	8.70%	184	transition	
2768	P	477	159	3	T -> C	83.70%	184	transition	
2780	P	489	163	3	G -> A	2.20%	184	transition	
2804	P	513	171	3	T -> C	3.80%	184	transition	
2837	P	546	182	3	G -> A	88.00%	184	transition	
2858	P	567	189	3	T -> C	2.20%	184	transition	
2873	P	582	194	3	T -> C	3.30%	184	transition	
2966	P	675	225	3	C -> T	2.20%	184	transition	
3011	P	720	240	3	T -> C	1.60%	184	transition	
3019					T -> C	1.60%	184	transition	
3054					T -> C	1.10%	184	transition	
3066					A -> G	1.10%	184	transition	
3085					T -> C	4.90%	184	transition	
3086					C -> T	2.20%	184	transition	
3091					C -> T	2.70%	184	transition	
3099					T -> C	5.40%	184	transition	
3103					C -> T	99.50%	184	transition	
3106					C -> T	88.00%	184	transition	
3109					C -> T	85.90%	184	transition	
3110					T -> C	2.20%	184	transition	
3123					A -> T	1.10%	184	transversion	
3135					G -> T/A	1.6%/1.1%	184	SNP	
3136					C -> T	2.20%	184	transition	

3142					T -> C	7.10%	184	transition	
3143					A -> T	1.60%	184	transversion	
3159					(A)6 -> (A)7	1.10%	184	Insertion (tandem repeat)	
3159					(A)6 -> (A)5	1.10%	184	Deletion (tandem repeat)	
3171					T -> C	99.50%	184	transition	
3181					(A)7 -> (A)8	1.60%	184	Insertion (tandem repeat)	
3217	M	18	6	3	C -> T	85.20%	183	transition	
3229	M	30	10	3	G -> A	8.70%	183	transition	
3247	M	48	16	3	T -> A	1.60%	183	transversion	
3266	M	67	23	1	C -> T	4.40%	183	transition	
3277	M	78	26	3	C -> T	1.10%	183	transition	
3310	M	111	37	3	C -> T	2.20%	183	transition	
3322	M	123	41	3	A -> G	2.20%	183	transition	
3339	M	140	47	2	T -> A	37.70%	183	transversion	L -> Q
3394	M	195	65	3	C -> T	1.10%	183	transition	
3397	M	198	66	3	A -> G	88.50%	183	transition	
3403	M	204	68	3	T -> C	88.50%	183	transition	
3409	M	210	70	3	G -> A	99.50%	183	transition	
3418	M	219	73	3	G -> A	1.60%	183	transition	M -> I
3466	M	267	89	3	C -> T	1.10%	183	transition	
3475	M	276	92	3	C -> T	33.90%	183	transition	
3505	M	306	102	3	G -> A	2.20%	183	transition	
3538	M	339	113	3	G -> A	1.10%	182	transition	
3548	M	349	117	1	C -> T	1.60%	182	transition	
3565	M	366	122	3	A -> T	4.40%	182	transversion	
3643	M	444	148	3	A -> G	1.60%	182	transition	
3670	M	471	157	3	A -> G	4.90%	182	transition	
3674	M	475	159	1	A -> G	1.10%	182	transition	I -> V
3697	M	498	166	3	T -> C	1.10%	183	transition	
3703	M	504	168	3	T -> C	88.50%	183	transition	
3721	M	522	174	3	G -> A	9.30%	183	transition	
3745	M	546	182	3	C -> T	1.10%	183	transition	
3763	M	564	188	3	A -> C	97.30%	182	transversion	
3763	M	564	188	3	A -> T	2.20%	182	transversion	
3787	M	588	196	3	T -> C	1.60%	182	transition	
3800	M	601	201	1	C -> T	1.10%	182	transition	
3823	M	624	208	3	C -> T	4.90%	182	transition	
3844	M	645	215	3	C -> T	1.10%	183	transition	
3969	M	770	257	2	A -> G	38.30%	183	transition	
3972					C -> T	4.40%	183	transition	
3976					T -> C	1.60%	183	transition	
3977					T -> C/-	2.2%/1.6%	183	Mixture	
3986					T -> A	10.40%	183	transversion	
3994					T -> C	82.00%	183	transition	
3998					T -> C	2.20%	183	transition	
4002					T -> C	82.00%	183	transition	
4009					T -> C	4.90%	183	transition	
4010					T -> A	1.10%	183	transversion	
4013					T -> C	82.50%	183	transition	
4014					A -> T	4.90%	183	transversion	
4017					T -> C	82.00%	183	transition	
4021					T -> C	2.20%	183	transition	
4030					T -> C	82.00%	183	transition	
4031					C -> T	1.10%	183	transition	
4039					T -> C	81.90%	182	transition	
4049					C -> T	6.00%	182	transition	
4053					C -> T	7.10%	182	transition	
4054					T -> C	81.90%	182	transition	
4057					G -> T	1.10%	182	transversion	
4064					T -> C	81.90%	182	transition	
4069					C -> T	11.50%	182	transition	
4093					A -> T	2.70%	182	transversion	
4107					C -> T	2.20%	182	transition	
4112					T -> A	1.60%	182	transversion	
4117					C -> T	3.80%	182	transition	

4118					G -> A	99.50%	182	transition	
4119					G -> A	4.40%	182	transition	
4123					C -> A	5.50%	182	transversion	
4129					T -> A	1.10%	182	transversion	
4131					A -> G	1.10%	182	transition	
4143					(A)6 -> (A)7	1.10%	184	Insertion (tandem repeat)	
4143					(A)6 -> (A)5	2.20%	182	Deletion (tandem repeat)	
4148					-AT	3.30%	182	Deletion	
4172					G -> A	1.10%	181	transition	
4175					G -> A	1.10%	181	transition	
4186					T -> C	1.70%	181	transition	
4189					T -> C	1.10%	181	transition	
4195					C -> G/A	36.5%/1.1%	181	SNP	
4228					T -> C	1.60%	182	transition	
4238					C -> T	1.10%	182	transition	
4287	SH	48	16	3	T -> C	1.10%	182	transition	
4296	SH	57	19	3	A -> G	1.10%	181	transition	
4297	SH	58	20	1	T -> C	99.50%	182	transition	
4300	SH	61	21	1	A -> G	1.10%	181	transition	I -> V
4300	SH	61	21	1	A -> T	1.10%	181	transversion	I -> L
4303	SH	64	22	1	C -> T	35.70%	182	transition	H -> Y
4335	SH	96	32	3	A -> C	99.40%	181	transversion	
4365	SH	126	42	3	C -> T	1.10%	181	transition	
4374	SH	135	45	3	C -> T	1.10%	181	transition	
4385	SH	146	49	2	T -> C	1.10%	181	transition	V -> A
4386	SH	147	49	3	A -> G	1.10%	181	transition	
4391	SH	152	51	2	A -> T	2.80%	181	transversion	H -> L
4397	SH	158	53	2	A -> G	2.80%	181	transition	K -> R
4398	SH	159	53	3	A -> G	1.10%	181	transition	
4423	SH	184	62	1	G -> A	5.50%	181	transition	V -> I
4459					A -> G	1.10%	182	transition	
4463					T -> C	3.30%	182	transition	
4480					G -> A	87.90%	182	transition	
4482					G -> A	1.60%	182	transition	
4495					C -> T	84.20%	183	transition	
4496					T -> C	3.30%	183	transition	
4520					T -> C	3.30%	183	transition	
4521					T -> C	99.50%	183	transition	
4532					C -> T	1.10%	183	transition	
4540					T -> C	3.30%	182	transition	
4543					A -> G	1.10%	182	transition	
4558					A -> T	3.30%	182	transversion	
4561					(A)5 -> (A)6	3.80%	184	Insertion (tandem repeat)	
4561					(A)5 -> (A)6	57.40%	183	Insertion (tandem repeat)	
4570					C -> T	1.10%	182	transition	
4575					C -> T	4.40%	182	transition	
4579					G -> A	1.10%	182	transition	
4585					G -> A	38.50%	182	transition	
4592					A -> G	2.20%	183	transition	
4598					(A)5 -> (A)6	4.90%	183	Insertion (tandem repeat)	
4606					C -> T	4.40%	183	transition	
4647	G	23	8	2	G -> A	1.60%	183	transition	R -> H
4651	G	27	9	3	C -> T	1.10%	183	transition	
4684	G	60	20	3	C -> T	2.20%	183	transition	
4691	G	67	23	1	CTA -> TTG	9.80%	183	Substitution	
4693	G	69	23	3	A -> G	71.60%	183	transition	
4708	G	84	28	3	G -> A	4.90%	183	transition	
4756	G	132	44	3	T -> C	1.10%	184	transition	
4763	G	139	47	1	G -> A	3.30%	184	transition	A -> T
4793	G	169	57	1	G -> A	32.20%	183	transition	A -> T
4842	G	218	73	2	C -> A	3.30%	183	transversion	T -> N

4864	G	240	80	3	G -> A	1.60%	183	transition	
4867	G	243	81	3	C -> T	1.10%	183	transition	
4885	G	261	87	3	C -> T	26.10%	184	transition	
4900	G	276	92	3	C -> T	2.70%	184	transition	
4907	G	283	95	1	C -> T	1.60%	183	transition	P -> S
4909	G	285	95	3	C -> T	1.10%	183	transition	
4927	G	303	101	3	C -> A	1.10%	183	transversion	F -> L
4945	G	321	107	3	T -> C	5.50%	183	transition	
4968	G	344	115	2	T -> C	6.60%	183	transition	L -> P
4969	G	345	115	3	A -> T	1.10%	183	transversion	
4977	G	353	118	2	C -> T	4.90%	183	transition	T -> I
4993	G	369	123	3	G -> A	1.10%	183	transition	
5007	G	383	128	2	C -> T	2.20%	182	transition	S -> F
5021	G	397	133	1	A -> G	1.10%	182	transition	I -> V
5031	G	407	136	2	C -> T	30.80%	182	transition	T -> I
5046	G	422	141	2	T -> C	1.10%	182	transition	I -> T
5048	G	424	142	1	CA -> TT	87.90%	182	Substitution	Q -> L
5053	G	429	143	3	T -> C	6.60%	182	transition	
5076	G	452	151	2	G -> A	1.10%	183	transition	R -> H
5081	G	457	153	1	A -> G	1.60%	183	transition	N -> D
5083	G	459	153	3	T -> C	1.10%	183	transition	
5095	G	471	157	3	C -> T	37.70%	183	transition	
5096	G	472	158	1	A -> C	1.60%	183	transversion	K -> Q
5104	G	480	160	3	C -> T	1.10%	183	transition	
5158	G	534	178	3	T -> C	97.30%	182	transition	
5221	G	597	199	3	C -> T	2.20%	182	transition	
5226	G	602	201	2	G -> A	98.90%	182	transition	R -> K
5230	G	606	202	3	C -> T	31.90%	182	transition	
5241	G	617	206	2	C -> A	37.40%	182	transversion	P -> Q
5244	G	620	207	2	C -> T	3.30%	182	transition	T -> I
5246	G	622	208	1	A -> C	88.50%	182	transversion	I -> L
5275	G	651	217	3	T -> C	4.90%	182	transition	
5277	G	653	218	2	A -> G	2.20%	182	transition	Q -> R
5297	G	673	225	1	G -> A	1.10%	182	transition	V -> I
5302	G	678	226	3	C -> A	1.60%	183	transversion	
5305	G	681	227	3	C -> T	79.80%	183	transition	
5308	G	684	228	3	C -> T	1.60%	183	transition	
5309	G	685	229	1	A -> G	1.10%	183	transition	K -> E
5314	G	690	230	3	C -> T	83.60%	183	transition	
5319	G	695	232	2	A -> G	84.20%	183	transition	E -> G
5333	G	709	237	1	G -> A	88.50%	183	transition	D -> N
5335	G	711	237	3	C -> T	1.10%	183	transition	
5338	G	714	238	3	C -> T	1.60%	183	transition	
5340	G	716	239	2	CC -> TT	1.10%	183	Substitution	T -> I
5341	G	717	239	3	C -> T	73.20%	183	transition	
5346	G	722	241	2	C -> T	1.10%	183	transition	T -> I
5355	G	731	244	2	G -> A	3.80%	183	transition	R -> K
5361	G	737	246	2	C -> T	1.10%	183	transition	T -> I
5363	G	739	247	1	C -> T	4.40%	183	transition	
5372	G	748	250	1	T -> C	5.50%	183	transition	S -> P
5373	G	749	250	2	C -> T	1.10%	183	transition	S -> F
5377	G	753	251	3	C -> A	1.60%	183	transversion	N -> K
5382	G	758	253	2	C -> A	84.20%	183	transversion	T -> K
5390	G	766	256	1	C -> T	1.10%	183	transition	P -> S
5396	G	772	258	1	C -> T	4.40%	183	transition	H -> Y
5411	G	787	263	1	G -> A	1.60%	183	transition	E -> K
5424	G	800	267	2	C -> T	2.70%	183	transition	S -> L
5435	G	811	271	1	G -> A	1.10%	183	transition	E -> K
5441	G	817	273	1	A -> C	7.10%	183	transversion	N -> H
5441	G	817	273	1	A -> T	81.40%	183	transversion	N -> Y
5444	G	820	274	1	CT -> TC	7.10%	183	Substitution	L -> S
5445	G	821	274	2	T -> C	77.00%	183	transition	L -> P
5462	G	838	280	1	C -> T	17.50%	183	transition	H -> Y
5462	G	838	280	1	CAT -> TAC	81.40%	183	Substitution	H -> Y
5472	G	848	283	2	C -> T	1.10%	183	transition	S -> F
5475	G	851	284	2	+GTCAAGAGGAAACCCTCCACTCAACCACCCCCGAAGGCCATCCAAGCCCATCACAAGTCCATACAACATCCG	2.20%	184	Insertion	E -> GQEETLHSTTPEGHPSPSQVHTTSE
5475	G	851	284	2	+GTCAAGAGGAAACCCTCCACTCAACCACCTCCGAAGGCTATCCAAGCCCATCACAAGTCCATACAACATCCG	76.60%	184	Insertion	E -> GQEETLHSTTSEGYPSPSQVHTTSE

5475	G	851	284	2	+GTCAAGAGGA AACCC TCCACTCAACC ACCTCC GAAGGCTATCT AAGCC CATCACAAGTC TATAC AACATCCG	2.20%	184	Insertion	E -> GQEETLHSTT SEG YLSPSQVYTT SE
5475	G	851	284	2	+TCCATCACAA GTCCA TACAACATCCG	1.10%	184	Insertion	
5481	G	857	286	2	C -> T	53.00%	183	transition	P -> L
5493	G	869	290	2	C -> T	82.50%	183	transition	P -> L
5493	G	869	290	2	CA -> TG	1.60%	183	Substitution	P -> L
5501	G	877	293	1	T -> A	1.10%	183	transversion	S -> T
5503	G	879	293	3	C -> A	6.00%	183	transversion	
5506	G	882	294	3	C -> T	1.10%	183	transition	
5509	G	885	295	3	A -> G	30.60%	183	transition	
5510	G	886	296	1	A -> C	3.80%	183	transversion	T -> P
5515	G	891	297	3	A -> C	3.30%	183	transversion	K -> N
5533	G	909	303	3	C -> T	4.40%	183	transition	
5543	G	919	307	1	C -> T	1.10%	183	transition	
5549	G	925	309	1	G -> A	1.10%	183	transition	A -> T
5551	G	927	309	3	C -> T	4.40%	183	transition	
5568	G	944	315	2	G -> A	1.10%	183	transition	R -> K
5572	G	948	316	3	C -> T	1.10%	183	transition	
5577	G	953	318	2	T -> C	3.80%	183	transition	I -> T
5583	G	959	320	2	C -> T	1.10%	183	transition	S -> F
5615	F	18	6	3	C -> T	7.20%	181	transition	
5633	F	36	12	3	C -> A	4.40%	181	transversion	
5633	F	36	12	3	C -> T	5.00%	181	transition	
5649	F	52	18	1	G -> A	1.10%	181	transition	V -> I
5663	F	66	22	3	C -> T	2.80%	181	transition	
5678	F	81	27	3	C -> T	1.70%	181	transition	
5720	F	123	41	3	C -> T	32.60%	181	transition	
5729	F	132	44	3	T -> C	1.60%	182	transition	
5783	F	186	62	3	T -> C	1.10%	180	transition	
5786	F	189	63	3	T -> C	4.40%	180	transition	
5816	F	219	73	3	C -> T	5.00%	181	transition	
5817	F	220	74	1	G -> A	5.50%	181	transition	A -> T
5819	F	222	74	3	T -> C	1.70%	181	transition	
5849	F	252	84	3	T -> C	1.70%	181	transition	
5857	F	260	87	2	A -> G	1.70%	181	transition	K -> R
5902	F	305	102	2	C -> T	1.70%	181	transition	A -> V
5906	F	309	103	3	C -> T	2.20%	181	transition	
5916	F	319	107	1	G -> A	1.10%	181	transition	A -> T
5928	F	331	111	1	C -> T	37.00%	181	transition	
5943	F	346	116	1	G -> A	88.50%	182	transition	D -> N
5961	F	364	122	1	A -> G	81.30%	182	transition	T -> A
5963	F	366	122	3	C -> T	1.60%	182	transition	
5983	F	386	129	2	T -> C	2.70%	182	transition	L -> S
5987	F	390	130	3	C -> T	88.50%	182	transition	
6041	F	444	148	3	C -> T	1.10%	182	transition	
6044	F	447	149	3	C -> T	5.50%	182	transition	
6050	F	453	151	3	C -> A	3.30%	181	transversion	
6059	F	462	154	3	A -> G	2.20%	182	transition	
6069	F	472	158	1	C -> T	1.10%	182	transition	
6074	F	477	159	3	C -> T	1.60%	182	transition	
6104	F	507	169	3	T -> C	1.10%	180	transition	
6111	F	514	172	1	C -> T	7.80%	180	transition	
6113	F	516	172	3	A -> G	2.20%	181	transition	
6116	F	519	173	3	C -> T	3.30%	181	transition	
6143	F	546	182	3	T -> C	5.00%	181	transition	
6170	F	573	191	3	A -> G	84.00%	181	transition	
6185	F	588	196	3	A -> G	1.70%	181	transition	
6213	F	616	206	1	A -> G	1.10%	181	transition	I -> V
6266	F	669	223	3	C -> T	1.10%	181	transition	
6275	F	678	226	3	G -> A	1.10%	181	transition	
6278	F	681	227	3	C -> T	1.10%	181	transition	
6287	F	690	230	3	A -> G	1.10%	182	transition	
6362	F	765	255	3	T -> C	1.10%	181	transition	
6374	F	777	259	3	A -> G	4.40%	181	transition	
6431	F	834	278	3	T -> C	1.10%	182	transition	
6476	F	879	293	3	A -> G	2.20%	182	transition	
6510	F	913	305	1	T -> C	99.50%	183	transition	
6530	F	933	311	3	T -> C	1.10%	182	transition	
6557	F	960	320	3	T -> C	5.00%	181	transition	
6578	F	981	327	3	G -> A	2.20%	182	transition	
6593	F	996	332	3	C -> T	1.10%	182	transition	
6641	F	1044	348	3	A -> T	6.60%	182	transversion	
6677	F	1080	360	3	T -> C	1.60%	182	transition	
6689	F	1092	364	3	G -> A	2.20%	182	transition	

6692	F	1095	365	3	A -> G	83.50%	182	transition	
6731	F	1134	378	3	G -> A	2.20%	182	transition	
6743	F	1146	382	3	T -> C	88.50%	182	transition	
6764	F	1167	389	3	C -> T	1.10%	184	transition	
6770	F	1173	391	3	T -> C	2.70%	184	transition	
6782	F	1185	395	3	T -> C	2.70%	184	transition	
6812	F	1215	405	3	C -> A	1.10%	183	transversion	
6812	F	1215	405	3	C -> T	2.20%	183	transition	
6825	F	1228	410	1	C -> T	2.20%	184	transition	
6830	F	1233	411	3	G -> A	99.50%	184	transition	
6845	F	1248	416	3	C -> T	3.80%	184	transition	
6851	F	1254	418	3	C -> A	1.10%	183	transversion	
6863	F	1266	422	3	T -> C	4.40%	183	transition	
6875	F	1278	426	3	T -> C	3.30%	183	transition	
6941	F	1344	448	3	T -> C	8.80%	182	transition	
6968	F	1371	457	3	T -> C	3.90%	181	transition	
6977	F	1380	460	3	T -> C	4.40%	181	transition	
7031	F	1434	478	3	T -> C	1.70%	181	transition	
7046	F	1449	483	3	C -> T	1.10%	180	transition	
7104	F	1507	503	1	C -> A	2.20%	181	transversion	L -> I
7109	F	1512	504	3	A -> G	32.00%	181	transition	
7181	F	1584	528	3	T -> C	2.80%	181	transition	
7184	F	1587	529	3	T -> C	89.00%	181	transition	
7199	F	1602	534	3	T -> A	5.50%	181	transversion	
7199	F	1602	534	3	T -> C	2.20%	181	transition	
7226	F	1629	543	3	T -> A	89.00%	181	transversion	
7229	F	1632	544	3	A -> T	89.00%	181	transversion	
7253	F	1656	552	3	C -> T	1.10%	181	transition	
7262	F	1665	555	3	A -> G	1.10%	181	transition	
7268	F	1671	557	3	C -> T	1.10%	181	transition	
7277	F	1680	560	3	T -> C	3.30%	181	transition	
7295	F	1698	566	3	T -> C	1.10%	181	transition	
7301	F	1704	568	3	T -> C	4.40%	181	transition	
7313	F	1716	572	3	T -> C	5.00%	181	transition	
7337					T -> C	7.70%	181	transition	
7340					T -> C	2.20%	181	transition	
7359					C -> T	1.10%	181	transition	
7360					A -> G	87.30%	181	transition	
7365					T -> A	4.40%	181	transversion	
7367					A -> G	98.90%	181	transition	
7374					G -> A	32.60%	181	transition	
7392					G -> A	84.50%	181	transition	
7393					G -> A	2.20%	181	transition	
7400					C -> T	3.90%	181	transition	
7405					A -> G	4.40%	181	transition	
7421					A -> T	1.10%	181	transversion	
7423					C -> T	1.10%	181	transition	
7437					C -> T	1.10%	182	transition	
7438					C -> T	3.80%	182	transition	
7453					A -> T	89.00%	181	transversion	
7457					A -> G	1.10%	181	transition	
7461					A -> G	82.20%	180	transition	
7470					T -> A	3.90%	180	transversion	
7484					-AA	87.40%	183	Deletion	
7489					C -> A	95.00%	180	transversion	
7491					A -> G	1.70%	180	transition	
7500					C -> T	6.70%	179	transition	
7569	M2-1	27	9	3	C -> T	2.20%	179	transition	
7611	M2-1	69	23	3	T -> C	3.30%	180	transition	
7650	M2-1	108	36	3	T -> A	2.70%	183	transversion	
7686	M2-1	144	48	3	G -> A	1.60%	182	transition	
7701	M2-1	159	53	3	C -> T	3.80%	182	transition	
7716	M2-1	174	58	3	A -> T	32.60%	181	transversion	
7771	M2-1	229	77	1	G -> A	1.10%	182	transition	V -> I
7782	M2-1	240	80	3	A -> T	1.10%	182	transversion	
7788	M2-1	246	82	3	T -> C	1.10%	182	transition	
7791	M2-1	249	83	3	T -> C	1.10%	182	transition	
7802	M2-1	260	87	2	T -> C	1.10%	182	transition	I -> T
7842	M2-1	300	100	3	C -> T	3.30%	182	transition	
7851	M2-1	309	103	3	C -> T	80.20%	182	transition	
7866	M2-1	324	108	3	C -> T	1.10%	182	transition	
7891	M2-1	349	117	1	C -> A	88.50%	183	transversion	H -> N

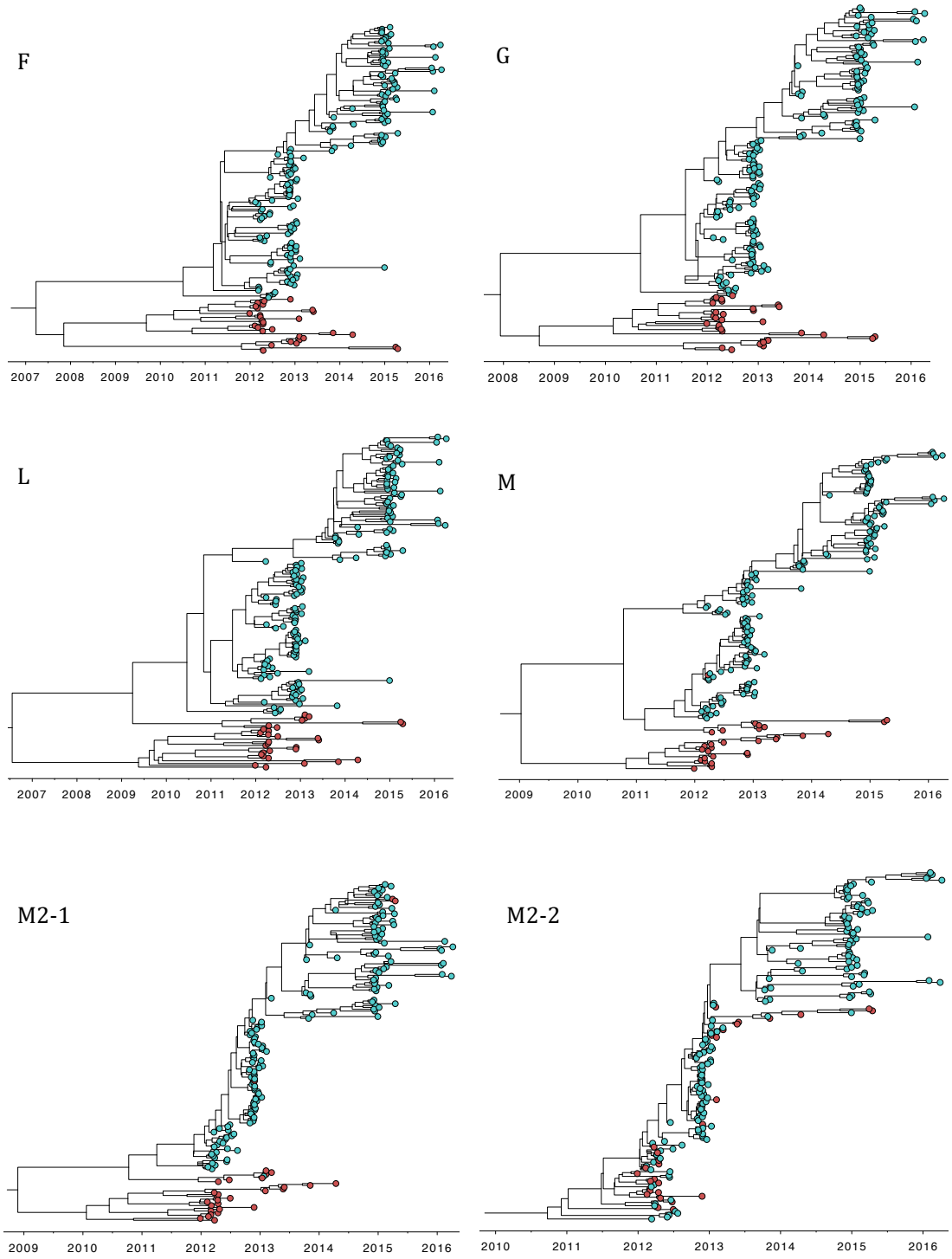
7899	M2-1	357	119	3	A -> G	99.50%	183	transition	
7901	M2-1	359	120	2	C -> T	1.10%	183	transition	P -> L
7902	M2-1	360	120	3	A -> C	1.60%	183	transversion	
7902	M2-1	360	120	3	A -> G	1.10%	183	transition	
7904	M2-1	362	121	2	G -> A	99.50%	183	transition	S -> N
7917	M2-1	375	125	3	A -> G	1.10%	184	transition	
7929	M2-1	387	129	3	T -> C	1.60%	184	transition	
7971	M2-1	429	143	3	A -> G	1.10%	184	transition	
7986	M2-1	444	148	3	G -> A	3.80%	184	transition	
8034	M2-1	492	164	3	A -> G	1.60%	184	transition	
8052	M2-1	510	170	3	C -> T	4.90%	184	transition	
8091	M2-1	549	183	3	T -> C	2.20%	184	transition	
8100	M2-1	558	186	3	C -> T	1.60%	184	transition	
8101	M2-1	559	187	1	C -> T	3.80%	184	transition	H -> Y
8105	M2-1	563	188	2	C -> T	1.10%	184	transition	A -> V
8105	M2-2	4	2	1	CCA -> TCT	1.10%	184	Substitution	P -> S
8107	M2-1	565	189	1	A -> T	1.10%	184	transversion	
8128	M2-2	27	9	3	C -> T	3.80%	184	transition	
8149	M2-2	48	16	3	T -> C	1.10%	182	transition	
8156	M2-2	55	19	1	C -> A	1.10%	182	transversion	L -> I
8166	M2-2	65	22	2	G -> A	2.20%	182	transition	S -> N
8167	M2-2	66	22	3	T -> C	1.10%	182	transition	
8185	M2-2	84	28	3	G -> A	2.70%	182	transition	M -> I
8215	M2-2	114	38	3	T -> C	1.10%	182	transition	
8232	M2-2	131	44	2	A -> G	2.20%	182	transition	N -> S
8246	M2-2	145	49	1	C -> T	2.20%	182	transition	P -> S
8251	M2-2	150	50	3	T -> A	1.10%	183	transversion	D -> E
8263	M2-2	162	54	3	T -> C	3.30%	183	transition	
8280	M2-2	179	60	2	C -> T	2.20%	183	transition	S -> F
8287	M2-2	186	62	3	C -> T	1.10%	182	transition	
8294	M2-2	193	65	1	G -> A	1.60%	182	transition	D -> N
8302	M2-2	201	67	3	T -> C	1.60%	182	transition	
8342	M2-2	241	81	1	T -> C	3.30%	182	transition	Y -> H
8359	M2-2	258	86	3	A -> G	37.90%	182	transition	
8378					C -> T	2.20%	182	transition	
8402					T -> C	1.60%	182	transition	
8421					T -> A	99.50%	182	transversion	
8459	L	25	9	1	T -> A	4.90%	182	transversion	S -> T
8521	L	87	29	3	T -> C	98.40%	182	transition	
8524	L	90	30	3	C -> T	84.60%	182	transition	
8536	L	102	34	3	C -> T	4.90%	183	transition	
8546	L	112	38	1	G -> A	1.60%	183	transition	G -> S
8557	L	123	41	3	C -> T	1.10%	183	transition	
8639	L	205	69	1	T -> A	35.70%	182	transversion	S -> T
8641	L	207	69	3	C -> T	37.40%	182	transition	
8653	L	219	73	3	G -> A	2.20%	183	transition	
8668	L	234	78	3	A -> G	1.60%	183	transition	
8731	L	297	99	3	G -> A	2.20%	182	transition	
8731	L	297	99	3	G -> T	2.70%	182	transversion	
8767	L	333	111	3	G -> A	2.80%	181	transition	
8791	L	357	119	3	C -> T	1.10%	181	transition	
8797	L	363	121	3	T -> C	2.20%	181	transition	
8867	L	433	145	1	C -> A	1.10%	183	transversion	Q -> K
8887	L	453	151	3	T -> C	84.10%	182	transition	
8890	L	456	152	3	T -> C	5.50%	182	transition	
8896	L	462	154	3	T -> C	99.50%	182	transition	
8962	L	528	176	3	T -> C	4.40%	182	transition	
8974	L	540	180	3	A -> C	2.20%	182	transversion	K -> N
8985	L	551	184	2	C -> T	9.30%	182	transition	T -> I
8989	L	555	185	3	C -> T	1.60%	182	transition	
9004	L	570	190	3	G -> A	9.30%	182	transition	
9034	L	600	200	3	T -> A	3.30%	183	transversion	
9073	L	639	213	3	A -> G	2.70%	183	transition	
9088	L	654	218	3	A -> G	1.10%	183	transition	
9094	L	660	220	3	G -> A	88.50%	183	transition	
9097	L	663	221	3	T -> C	1.10%	183	transition	
9133	L	699	233	3	G -> A	1.10%	184	transition	
9208	L	774	258	3	C -> T	2.20%	183	transition	
9247	L	813	271	3	A -> G	2.70%	182	transition	
9274	L	840	280	3	A -> G	6.60%	183	transition	
9277	L	843	281	3	T -> C	1.10%	183	transition	
9313	L	879	293	3	C -> T	6.50%	184	transition	
9364	L	930	310	3	G -> A	1.10%	183	transition	
9377	L	943	315	1	C -> A	1.10%	184	transversion	L -> I
9403	L	969	323	3	A -> G	1.10%	184	transition	
9409	L	975	325	3	C -> T	1.10%	184	transition	
9436	L	1002	334	3	G -> A	2.70%	184	transition	
9437	L	1003	335	1	G -> A	1.10%	184	transition	V -> I
9442	L	1008	336	3	G -> A	4.30%	184	transition	
9457	L	1023	341	3	T -> C	37.20%	183	transition	

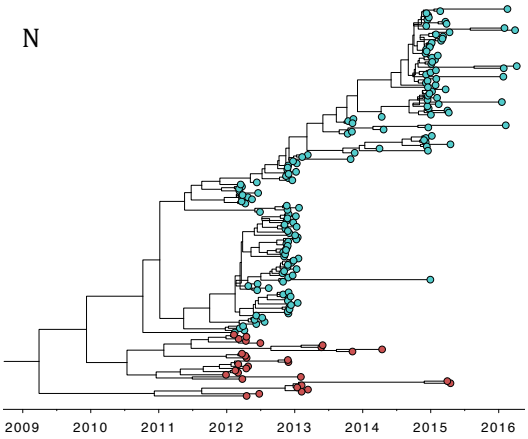
9517	L	1083	361	3	C -> T	3.80%	183	transition	
9535	L	1101	367	3	C -> T	2.20%	183	transition	
9538	L	1104	368	3	T -> A	4.40%	183	transversion	
9538	L	1104	368	3	T -> C	83.60%	183	transition	
9541	L	1107	369	3	T -> C	83.60%	183	transition	
9550	L	1116	372	3	G -> A	88.00%	183	transition	
9559	L	1125	375	3	G -> T	1.10%	183	transversion	
9610	L	1176	392	3	A -> C	2.20%	183	transversion	
9644	L	1210	404	1	T -> C	1.10%	184	transition	F -> L
9646	L	1212	404	3	C -> T	3.80%	182	transition	
9649	L	1215	405	3	T -> A	88.00%	184	transversion	
9689	L	1255	419	1	C -> T	1.10%	184	transition	
9715	L	1281	427	3	A -> G	8.20%	184	transition	
9751	L	1317	439	3	C -> T	2.20%	184	transition	
9766	L	1332	444	3	G -> A	98.40%	184	transition	
9796	L	1362	454	3	G -> A	1.10%	184	transition	
9859	L	1425	475	3	T -> C	1.60%	183	transition	
9898	L	1464	488	3	T -> C	84.20%	183	transition	
9933	L	1499	500	2	A -> G	1.10%	183	transition	N -> S
9955	L	1521	507	3	A -> G	4.90%	183	transition	
9991	L	1557	519	3	A -> G	4.90%	183	transition	
10015	L	1581	527	3	G -> A	2.20%	183	transition	
10018	L	1584	528	3	C -> T	99.50%	183	transition	
10033	L	1599	533	3	T -> C	24.00%	183	transition	
10066	L	1632	544	3	T -> A	1.10%	183	transversion	
10111	L	1677	559	3	G -> A	3.30%	183	transition	
10150	L	1716	572	3	A -> G	10.40%	182	transition	
10177	L	1743	581	3	A -> G	1.10%	182	transition	
10207	L	1773	591	3	A -> G	1.10%	182	transition	
10216	L	1782	594	3	A -> G	2.20%	182	transition	
10226	L	1792	598	1	T -> C	73.80%	183	transition	Y -> H
10226	L	1792	598	1	TAC -> CAT	10.40%	183	Substitution	Y -> H
10229	L	1795	599	1	A -> G	1.60%	183	transition	N -> D
10230	L	1796	599	2	A -> C	3.80%	183	transversion	N -> T
10261	L	1827	609	3	T -> C	99.50%	184	transition	
10264	L	1830	610	3	G -> T	1.10%	184	transversion	
10280	L	1846	616	1	T -> C	1.10%	184	transition	
10337	L	1903	635	1	A -> G	1.10%	184	transition	M -> V
10360	L	1926	642	3	A -> G	1.10%	184	transition	
10378	L	1944	648	3	T -> A	88.00%	184	transversion	
10384	L	1950	650	3	C -> T	2.20%	184	transition	
10522	L	2088	696	3	T -> C	1.10%	184	transition	
10585	L	2151	717	3	C -> T	88.50%	182	transition	
10591	L	2157	719	3	A -> G	33.00%	182	transition	
10627	L	2193	731	3	C -> T	1.60%	183	transition	
10639	L	2205	735	3	A -> G	10.40%	183	transition	
10645	L	2211	737	3	T -> A	3.90%	181	transversion	
10700	L	2266	756	1	C -> T	2.20%	182	transition	H -> Y
10705	L	2271	757	3	T -> C	3.30%	182	transition	
10739	L	2305	769	1	T -> C	84.60%	182	transition	
10744	L	2310	770	3	T -> C	1.10%	182	transition	
10804	L	2370	790	3	C -> A	99.40%	180	transversion	
10808	L	2374	792	1	C -> T	1.70%	180	transition	
10810	L	2376	792	3	A -> G	3.90%	180	transition	
10858	L	2424	808	3	T -> C	1.10%	181	transition	
10909	L	2475	825	3	A -> G	1.60%	182	transition	
10912	L	2478	826	3	T -> C	1.10%	182	transition	
10927	L	2493	831	3	A -> G	1.10%	182	transition	
10957	L	2523	841	3	T -> C	89.00%	182	transition	
10966	L	2532	844	3	G -> A	1.60%	182	transition	
10972	L	2538	846	3	A -> G	1.60%	182	transition	
11020	L	2586	862	3	C -> A	88.00%	183	transversion	
11038	L	2604	868	3	T -> C	8.70%	183	transition	
11050	L	2616	872	3	G -> A	9.80%	183	transition	
11062	L	2628	876	3	C -> T	1.10%	183	transition	
11068	L	2634	878	3	A -> G	3.80%	183	transition	
11107	L	2673	891	3	A -> G	1.10%	183	transition	
11143	L	2709	903	3	G -> A	1.10%	181	transition	
11176	L	2742	914	3	G -> A	99.40%	181	transition	
11248	L	2814	938	3	A -> G	1.10%	178	transition	
11261	L	2827	943	1	A -> C	3.40%	178	transversion	N -> H
11296	L	2862	954	3	C -> T	89.30%	178	transition	
11317	L	2883	961	3	C -> T	2.20%	178	transition	
11323	L	2889	963	3	A -> G	1.10%	178	transition	
11375	L	2941	981	1	T -> C	3.90%	181	transition	
11509	L	3075	1025	3	A -> G	1.10%	181	transition	
11521	L	3087	1029	3	A -> G	1.10%	181	transition	
11528	L	3094	1032	1	T -> G	5.00%	181	transversion	S -> A
11560	L	3126	1042	3	C -> T	1.10%	181	transition	
11569	L	3135	1045	3	A -> G	97.80%	181	transition	
11575	L	3141	1047	3	C -> T	38.10%	181	transition	
11581	L	3147	1049	3	C -> T	6.60%	181	transition	

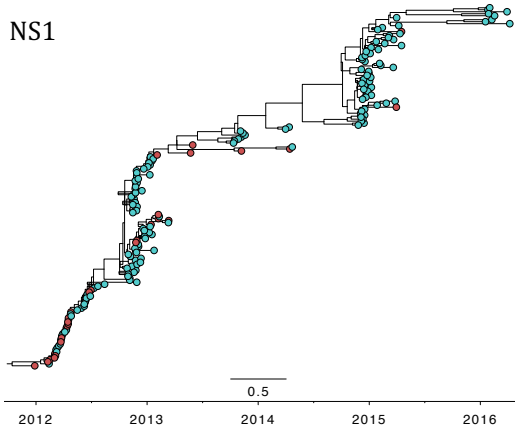
11596	L	3162	1054	3	C -> T	1.70%	181	transition	
11713	L	3279	1093	3	C -> T	89.00%	182	transition	
11752	L	3318	1106	3	C -> T	89.00%	181	transition	
11798	L	3364	1122	1	C -> T	1.10%	181	transition	
11821	L	3387	1129	3	A -> G	1.70%	181	transition	
11854	L	3420	1140	3	T -> C	1.10%	180	transition	
11929	L	3495	1165	3	T -> C	88.50%	182	transition	
11991	L	3557	1186	2	G -> A	1.10%	181	transition	R -> K
12005	L	3571	1191	1	A -> G	1.10%	182	transition	I -> V
12013	L	3579	1193	3	T -> C	1.10%	182	transition	
12025	L	3591	1197	3	A -> G	1.10%	182	transition	
12076	L	3642	1214	3	C -> T	6.60%	182	transition	
12142	L	3708	1236	3	T -> C	1.60%	182	transition	
12256	L	3822	1274	3	A -> T	1.60%	182	transversion	
12350	L	3916	1306	1	G -> A	3.30%	184	transition	D -> N
12364	L	3930	1310	3	G -> A	1.10%	184	transition	
12418	L	3984	1328	3	T -> C	88.00%	184	transition	
12454	L	4020	1340	3	T -> A	3.30%	184	transversion	
12487	L	4053	1351	3	A -> T	1.10%	184	transversion	
12544	L	4110	1370	3	C -> T	1.60%	184	transition	
12556	L	4122	1374	3	A -> G	81.50%	184	transition	
12625	L	4191	1397	3	A -> T	83.20%	184	transversion	
12637	L	4203	1401	3	A -> G	32.60%	184	transition	
12648	L	4214	1405	2	T -> C	1.10%	184	transition	V -> A
12667	L	4233	1411	3	A -> T	87.50%	184	transversion	
12676	L	4242	1414	3	C -> T	3.80%	184	transition	
12679	L	4245	1415	3	G -> A	35.30%	184	transition	
12697	L	4263	1421	3	G -> A	1.10%	184	transition	
12710	L	4276	1426	1	A -> G	1.10%	184	transition	I -> V
12793	L	4359	1453	3	G -> A	2.70%	184	transition	
12799	L	4365	1455	3	A -> G	37.00%	184	transition	
12870	L	4436	1479	2	C -> T	1.10%	184	transition	A -> V
12871	L	4437	1479	3	G -> A	2.20%	184	transition	
12895	L	4461	1487	3	T -> C	1.10%	184	transition	
12898	L	4464	1488	3	T -> C	1.10%	184	transition	
12905	L	4471	1491	1	A -> G	1.10%	184	transition	I -> V
12907	L	4473	1491	3	T -> C	1.10%	184	transition	
12922	L	4488	1496	3	A -> G	1.10%	184	transition	
13033	L	4599	1533	3	T -> C	1.10%	181	transition	
13060	L	4626	1542	3	C -> A	1.60%	183	transversion	
13066	L	4632	1544	3	T -> C	2.20%	183	transition	
13069	L	4635	1545	3	T -> C	8.80%	182	transition	
13072	L	4638	1546	3	T -> C	2.20%	183	transition	
13082	L	4648	1550	1	G -> A	2.20%	183	transition	G -> S
13117	L	4683	1561	3	A -> T	1.10%	184	transversion	
13162	L	4728	1576	3	G -> A	1.10%	182	transition	
13180	L	4746	1582	3	T -> C	7.10%	183	transition	
13210	L	4776	1592	3	A -> C	8.70%	183	transversion	
13210	L	4776	1592	3	A -> T	90.70%	183	transversion	
13213	L	4779	1593	3	C -> T	5.50%	183	transition	
13223	L	4789	1597	1	A -> T	1.10%	183	transversion	S -> C
13315	L	4881	1627	3	T -> A	1.10%	183	transversion	
13344	L	4910	1637	2	A -> G	1.10%	181	transition	K -> R
13348	L	4914	1638	3	A -> T	1.60%	182	transversion	
13369	L	4935	1645	3	A -> T	99.40%	180	transversion	
13435	L	5001	1667	3	A -> G	88.80%	179	transition	
13453	L	5019	1673	3	T -> G	88.80%	179	transversion	
13456	L	5022	1674	3	T -> C	1.10%	179	transition	
13519	L	5085	1695	3	T -> C	2.20%	179	transition	
13522	L	5088	1696	3	T -> C	88.80%	179	transition	
13558	L	5124	1708	3	T -> C	88.90%	180	transition	
13570	L	5136	1712	3	C -> T	2.20%	180	transition	
13579	L	5145	1715	3	C -> T	99.40%	180	transition	
13582	L	5148	1716	3	A -> G	2.20%	180	transition	I -> M
13608	L	5174	1725	2	AT -> GA	6.70%	180	Substitution	D -> G
13609	L	5175	1725	3	T -> A	82.20%	180	transversion	D -> E
13609	L	5175	1725	3	T -> C	4.40%	180	transition	
13611	L	5177	1726	2	A -> G	39.80%	181	transition	K -> R
13615	L	5181	1727	3	G -> A	1.10%	181	transition	
13623	L	5189	1730	2	G -> A	1.10%	180	transition	S -> N
13639	L	5205	1735	3	T -> C	4.90%	182	transition	
13646	L	5212	1738	1	G -> A	3.80%	182	transition	V -> I
13648	L	5214	1738	3	T -> A	2.20%	182	transversion	
13651	L	5217	1739	3	C -> T	1.10%	182	transition	
13664	L	5230	1744	1	C -> T	2.70%	182	transition	P -> S
13674	L	5240	1747	2	C -> T	1.10%	182	transition	S -> F
13681	L	5247	1749	3	G -> A	1.10%	182	transition	
13695	L	5261	1754	2	C -> T	1.60%	182	transition	S -> L
13720	L	5286	1762	3	C -> T	2.70%	182	transition	
13725	L	5291	1764	2	G -> A	1.10%	181	transition	R -> K
13736	L	5302	1768	1	T -> C	1.10%	182	transition	Y -> H
13753	L	5319	1773	3	A -> G	86.80%	182	transition	

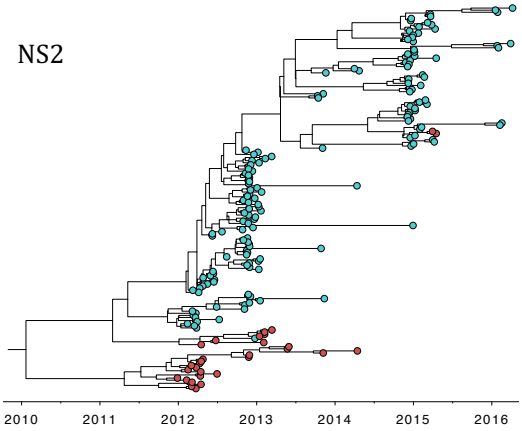
13792	L	5358	1786	3	A -> T	5.50%	182	transversion	
13795	L	5361	1787	3	C -> T	1.10%	182	transition	
13837	L	5403	1801	3	C -> A	4.90%	182	transversion	
13891	L	5457	1819	3	T -> C	2.20%	182	transition	
13906	L	5472	1824	3	A -> G	1.10%	181	transition	
13960	L	5526	1842	3	A -> G	35.90%	181	transition	
13969	L	5535	1845	3	T -> A	35.90%	181	transversion	D -> E
13981	L	5547	1849	3	A -> G	1.10%	181	transition	I -> M
13993	L	5559	1853	3	T -> C	3.80%	183	transition	
14122	L	5688	1896	3	G -> A	88.00%	184	transition	
14179	L	5745	1915	3	C -> T	1.60%	183	transition	
14251	L	5817	1939	3	G -> A	3.80%	183	transition	
14261	L	5827	1943	1	G -> A	2.20%	183	transition	V -> I
14302	L	5868	1956	3	A -> G	1.10%	184	transition	
14311	L	5877	1959	3	C -> T	37.50%	184	transition	
14374	L	5940	1980	3	C -> T	1.10%	184	transition	
14425	L	5991	1997	3	C -> T	2.20%	184	transition	
14491	L	6057	2019	3	A -> G	4.30%	184	transition	
14506	L	6072	2024	3	A -> G	1.10%	183	transition	
14522	L	6088	2030	1	C -> T	1.10%	183	transition	
14614	L	6180	2060	3	C -> T	1.10%	183	transition	
14629	L	6195	2065	3	A -> G	2.20%	183	transition	
14656	L	6222	2074	3	T -> A	99.50%	182	transversion	F -> L
14665	L	6231	2077	3	C -> T	99.50%	182	transition	
14669	L	6235	2079	1	A -> G	1.10%	182	transition	S -> G
14675	L	6241	2081	1	G -> A	1.10%	182	transition	D -> N
14686	L	6252	2084	3	G -> A	89.30%	182	transition	
14757	L	6323	2108	2	A -> G	1.10%	182	transition	K -> R
14797	L	6363	2121	3	A -> T	3.90%	181	transversion	L -> F
14803	L	6369	2123	3	T -> C	3.90%	181	transition	
14806	L	6372	2124	3	T -> C	3.90%	181	transition	
14819	L	6385	2129	1	G -> A	1.10%	180	transition	V -> I
14884	L	6450	2150	3	A -> G	1.10%	179	transition	

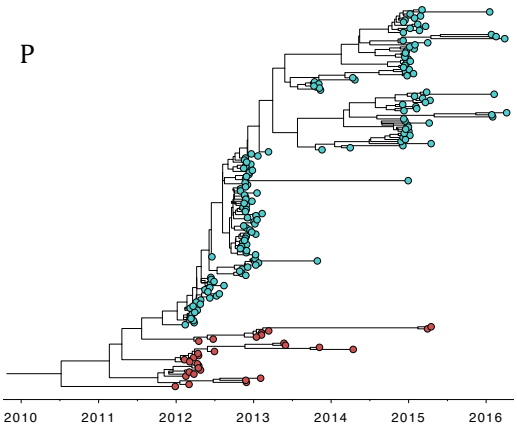
7.7 BEAST MCC trees showing divergence between ON1 (cyan) and GA2 (red) viruses across different RSV ORFs using Kilifi RSV-A dataset

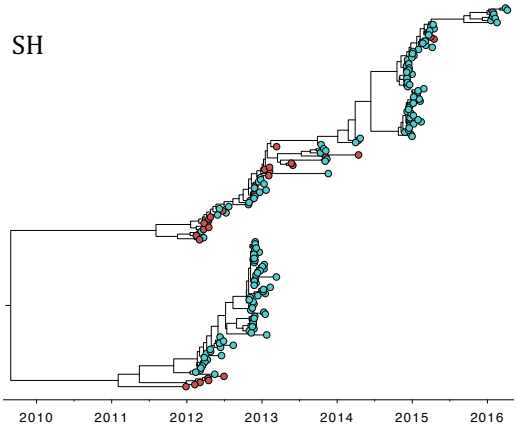


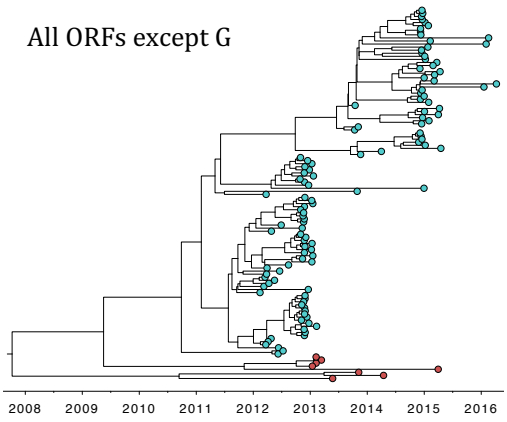

N
2009
2010
2011
2012
2013
2014
2015
2016

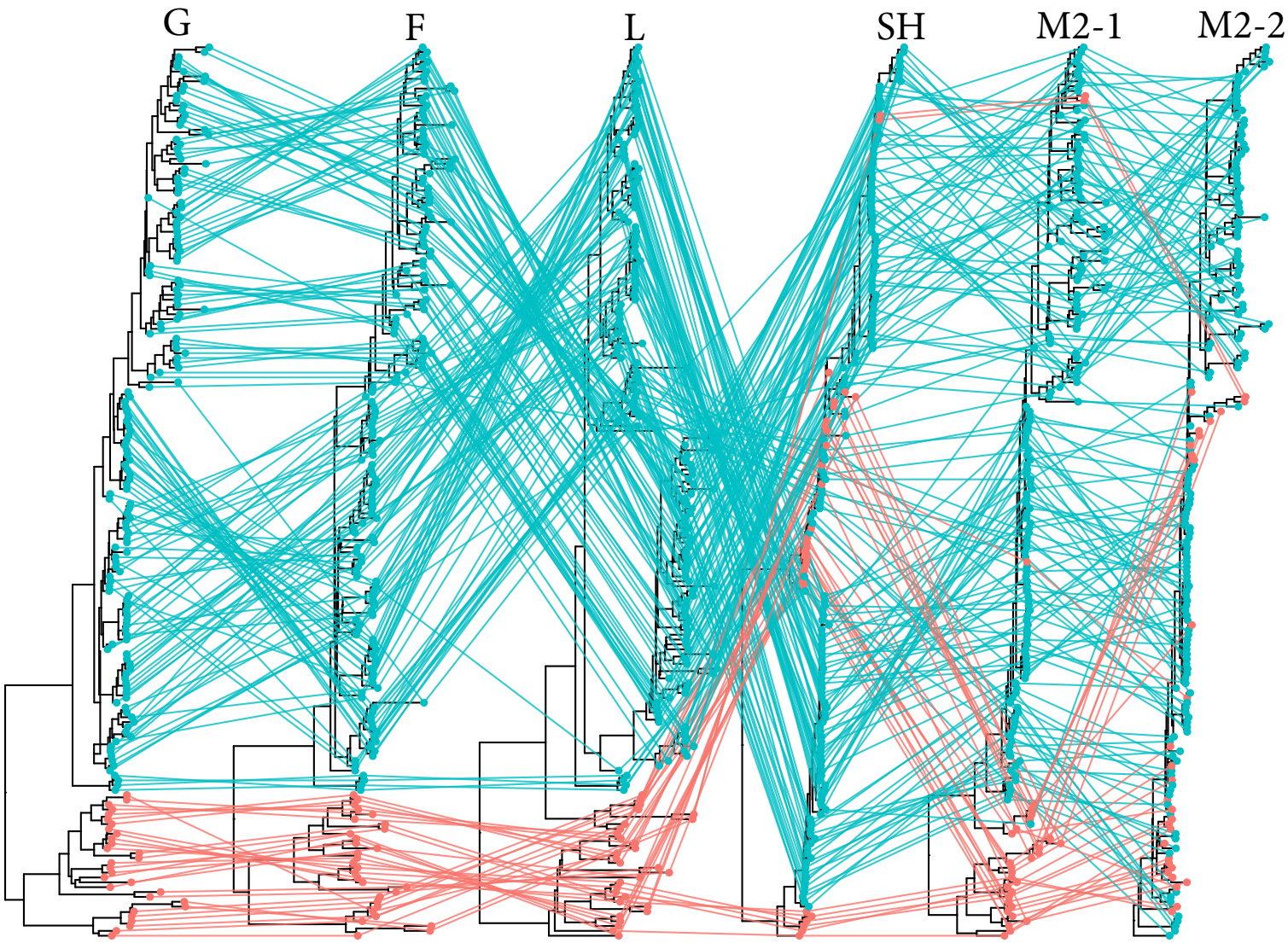

NS1
0.5
2012
2013
2014
2015
2016

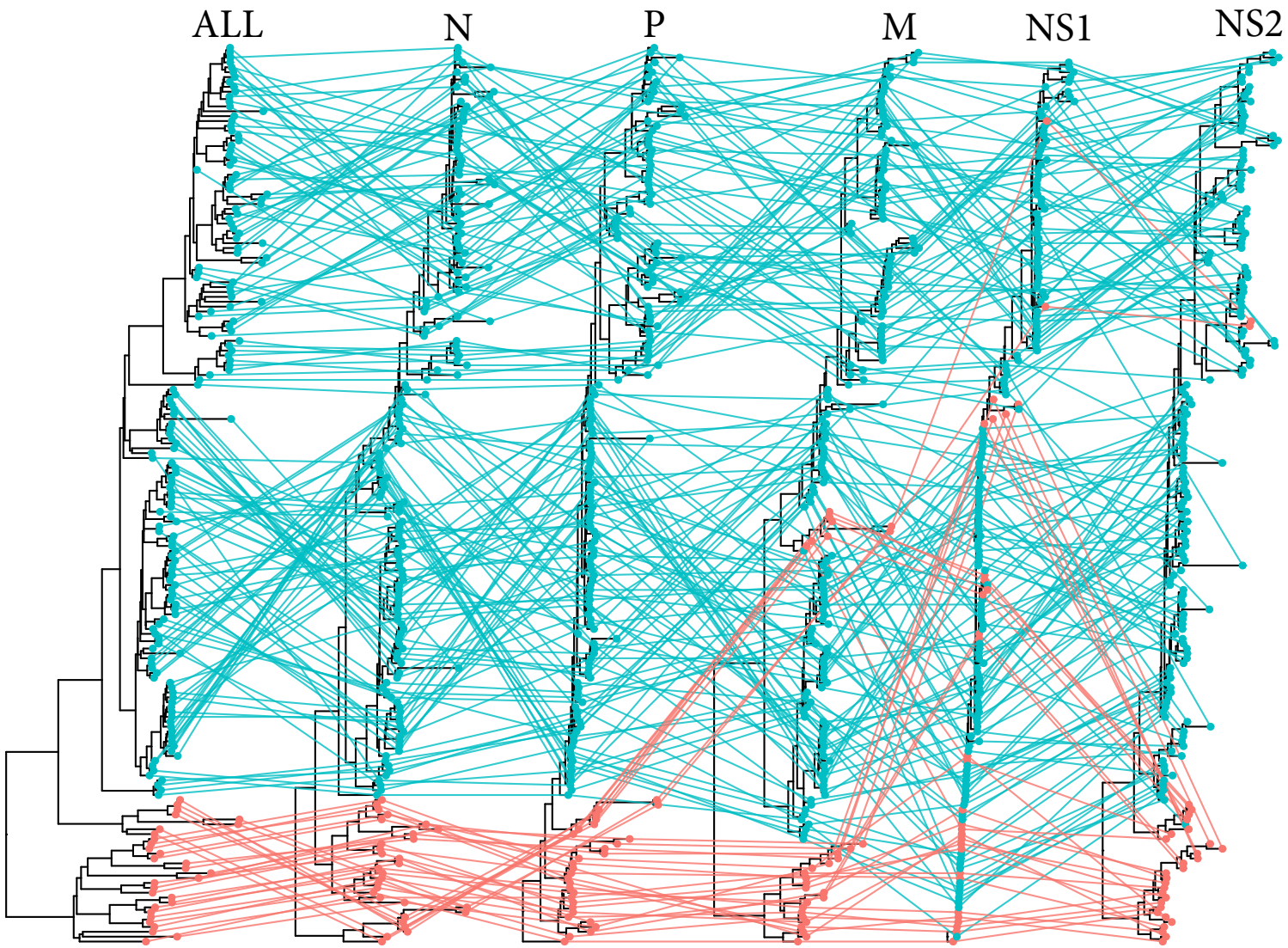

NS2
2010
2011
2012
2013
2014
2015
2016


P
2010
2011
2012
2013
2014
2015
2016


SH
2010
2011
2012
2013
2014
2015
2016


All ORFs except G
2008
2009
2010
2011
2012
2013
2014
2015
2016


G
F
L
SH
M2-1
M2-2


ALL
N
P
M
NS1
NS2

7.8 Signature SNPs between successful and limited transmission ON1 viruses

CDS	CDS Nt Pos.	SNP Codon Pos.	CDS AA Pos.	Change	AA Change	SNP Type
N	151	1	51	T -> C		Transition
N	558	3	186	T -> C		Transition
N	561	3	187	T -> C		Transition
N	717	3	239	T -> C		Transition
P	567	3	189	T -> C		Transition
M	111	3	37	C -> T		Transition
M	123	3	41	A -> G		Transition
G	69	3	23	G -> A		Transition
G	383	2	128	C -> T	S -> F	Transition
G	821	2	274	C -> T	P -> L	Transition
G	840	3	280	C -> T		Transition
G	893	2	298	C -> T	P -> L	Transition
G	910	1	304	C -> T	H -> Y	Transition
G	929	2	310	C -> T	P -> L	Transition
F	309	3	103	C -> T		Transition
F	364	1	122	G -> A	A -> T	Transition
F	516	3	172	A -> G		Transition
F	1507	1	503	C -> A	L -> I	Transversion
M2-1	549	3	183	T -> C		Transition
M2-2	131	2	44	A -> G	N -> S	Transition
L	219	3	73	G -> A		Transition
L	297	3	99	G -> T		Transversion
L	528	3	176	T -> C		Transition
L	540	3	180	A -> C	K -> N	Transversion
L	774	3	258	C -> T		Transition
L	1002	3	334	G -> A		Transition
L	1101	3	367	C -> T		Transition
L	1599	3	533	T -> C		Transition
L	1782	3	594	A -> G		Transition
L	4122	3	1374	G -> A		Transition
L	4437	3	1479	G -> A		Transition
L	5457	3	1819	T -> C		Transition
L	5991	3	1997	C -> T		Transition

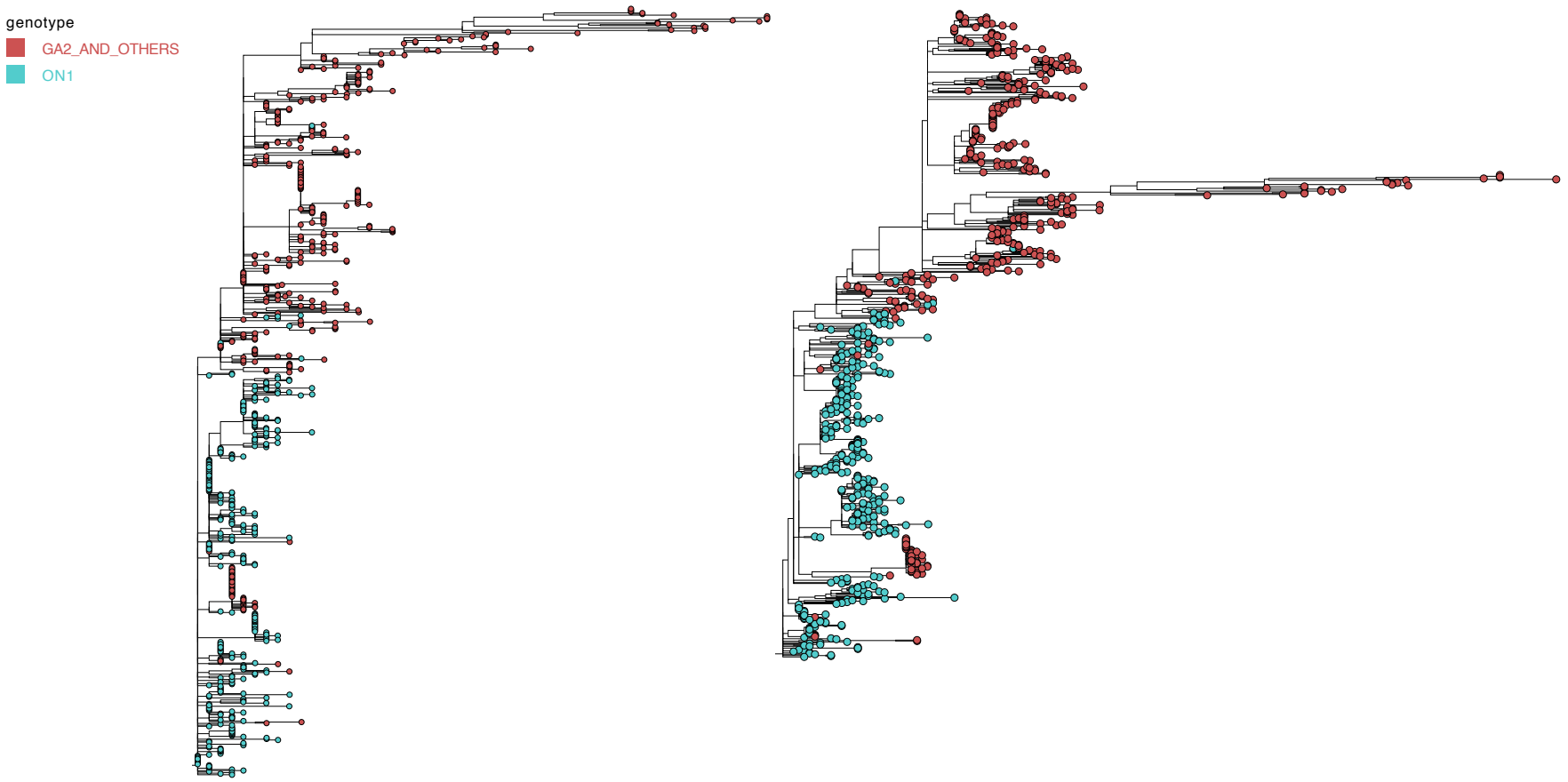
7.9 Signature SNPs between non-BA and BA viruses

ORF	ORF Nt Pos.	ORF AA Pos.	Change	AA Change	SNP Type
NS1 CDS	133	45	G -> A	A -> T	Substitution
NS1 CDS	315	105	G -> A	M -> I	Substitution
NS1 CDS	412	138	A -> C	N -> H	Substitution
N CDS	1114	372	G -> A	A -> T	Substitution
M CDS	266	89	T -> C	I -> T	Substitution
SH CDS	170	57	T -> A	L -> Q	Substitution
G CDS	10	4	C -> A	H -> N	Substitution
G CDS	95	32	G -> A	R -> K	Substitution
G CDS	301	101	T -> C	S -> P	Substitution
G CDS	325	109	T -> C	S -> P	Substitution
G CDS	412	138	ACC -> TCT	T -> S	Substitution
G CDS	418	140	T -> C	S -> P	Substitution
G CDS	428	143	C -> A	T -> N	Substitution
G CDS	472	158	-AAA	K ->	Deletion
G CDS	619	207	A -> C	T -> P	Substitution
G CDS	656	219	C -> T	P -> L	Substitution
G CDS	686	229	CC -> TT	T -> I	Substitution
G CDS	710	237	T -> C	L -> P	Substitution
G CDS	739	247	T -> C	S -> P	Substitution
G CDS	757	253	60 nt insertion	30 AA insertion	Insertion
G CDS	799	267	C -> T	H -> Y	Substitution
G CDS	883	295	AC -> CT	T -> L	Substitution
G CDS	892	298	C -> T	H -> Y	Substitution
F CDS	135	45	T -> G	F -> L	Substitution
F CDS	1585	529	A -> G	T -> A	Substitution
M2-1 CDS	425	142	A -> G	N -> S	Substitution
M2-1 CDS	545	182	A -> G	N -> S	Substitution
M2-2 CDS	103	35	A -> G	N -> D	Substitution
L CDS	23	8	A -> G	N -> S	Substitution
L CDS	529	177	C -> T	H -> Y	Substitution
L CDS	551	184	C -> A	T -> N	Substitution
L CDS	1120	374	A -> G	N -> D	Substitution
L CDS	2918	973	T -> C	M -> T	Substitution
L CDS	3748	1250	A -> G	S -> G	Substitution
L CDS	4640	1547	A -> G	K -> R	Substitution
L CDS	5148	1716	G -> A	M -> I	Substitution
L CDS	5178	1726	T -> A	S -> R	Substitution
L CDS	5291	1764	A -> G	K -> R	Substitution
L CDS	5360	1787	C -> A	A -> E	Substitution
L CDS	6125	2042	C -> T	T -> I	Substitution
L CDS	6195	2065	A -> T	K -> N	Substitution

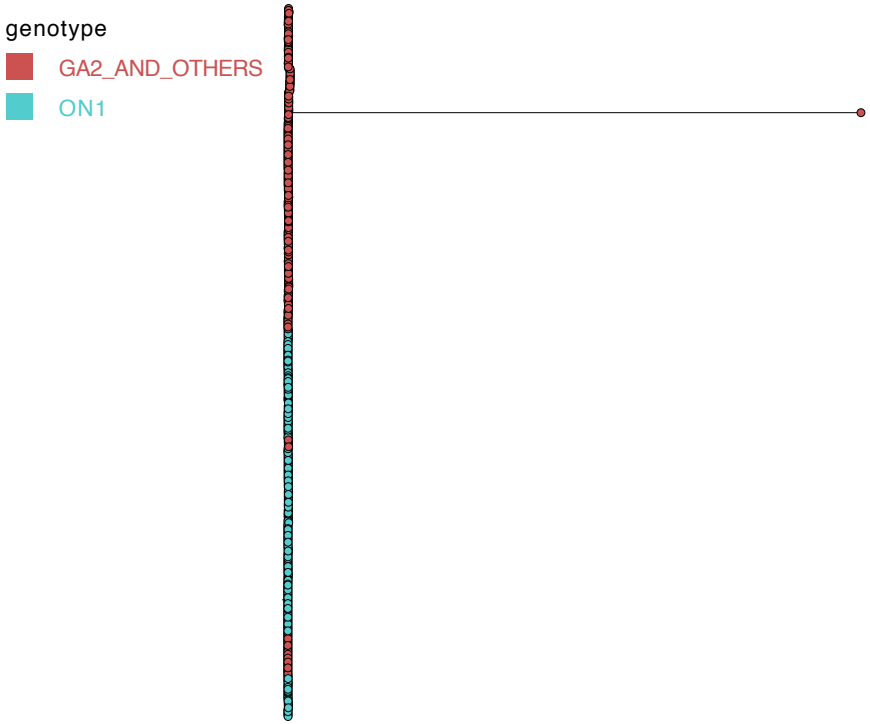
7.10 ML trees showing clustering between ON1 (cyan) and non-ON1 (red) viruses using global RSV-A dataset

RSV-A F-gene

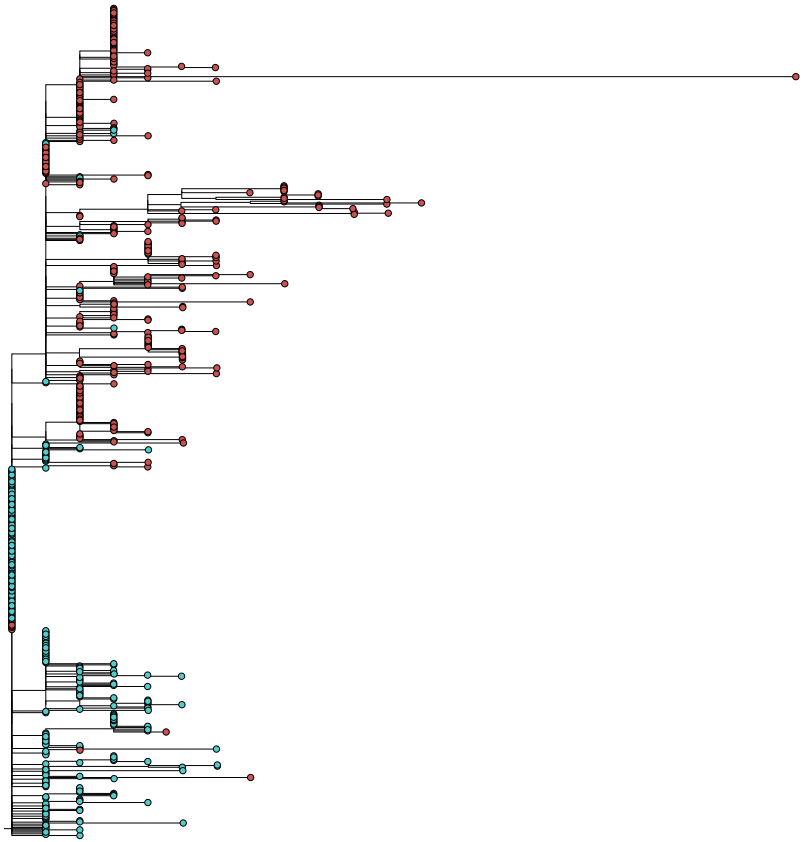
RSV-A L-gene



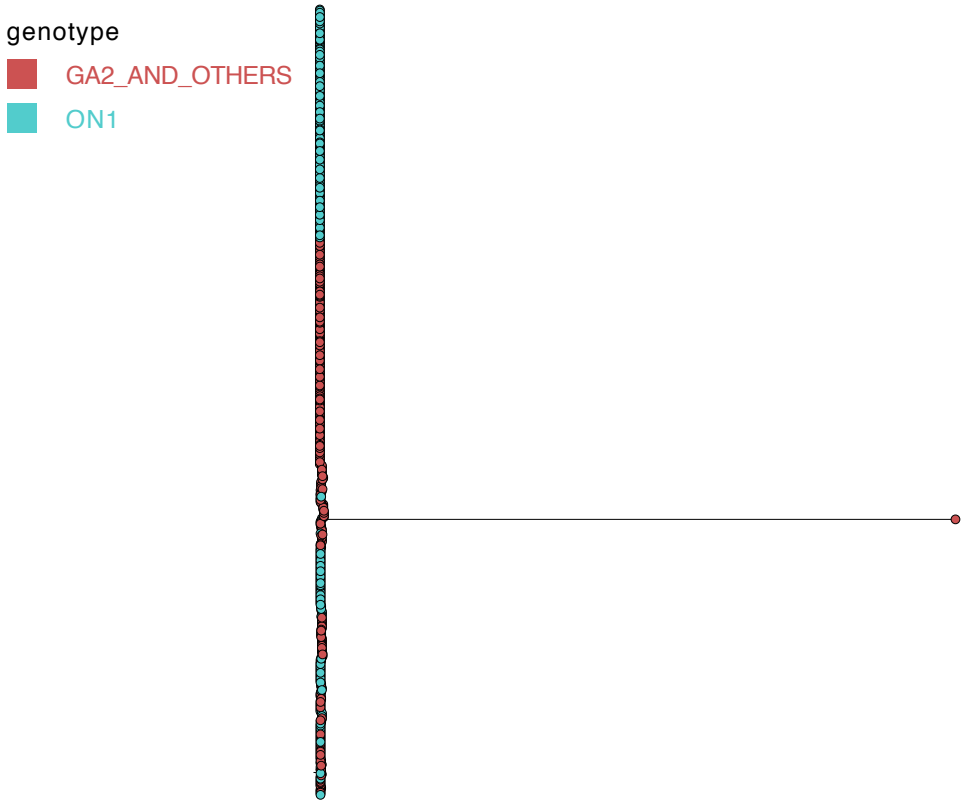
RSV-A N-gene



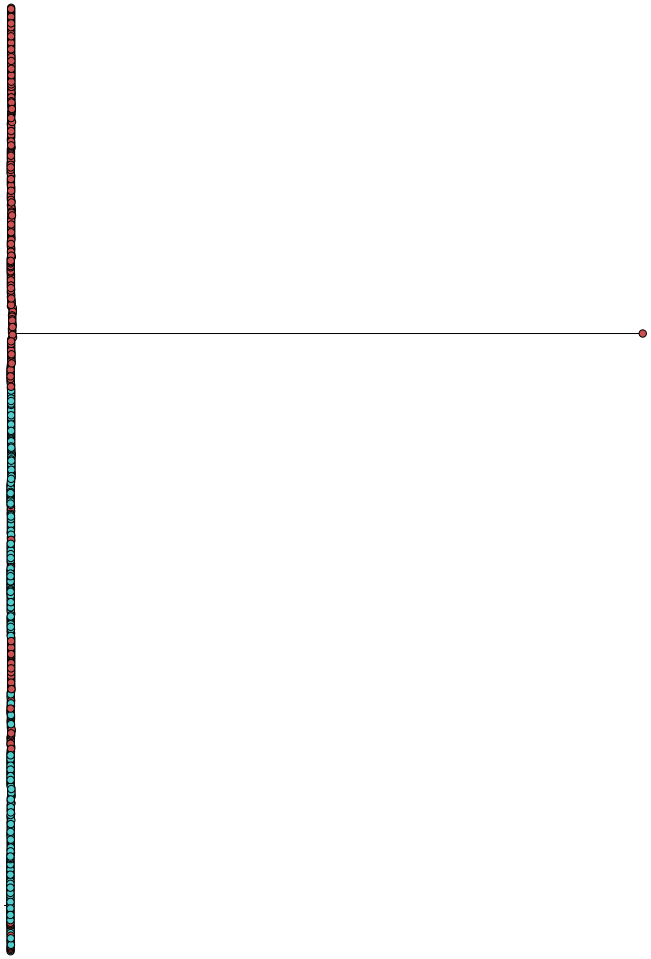
RSV-A P-gene



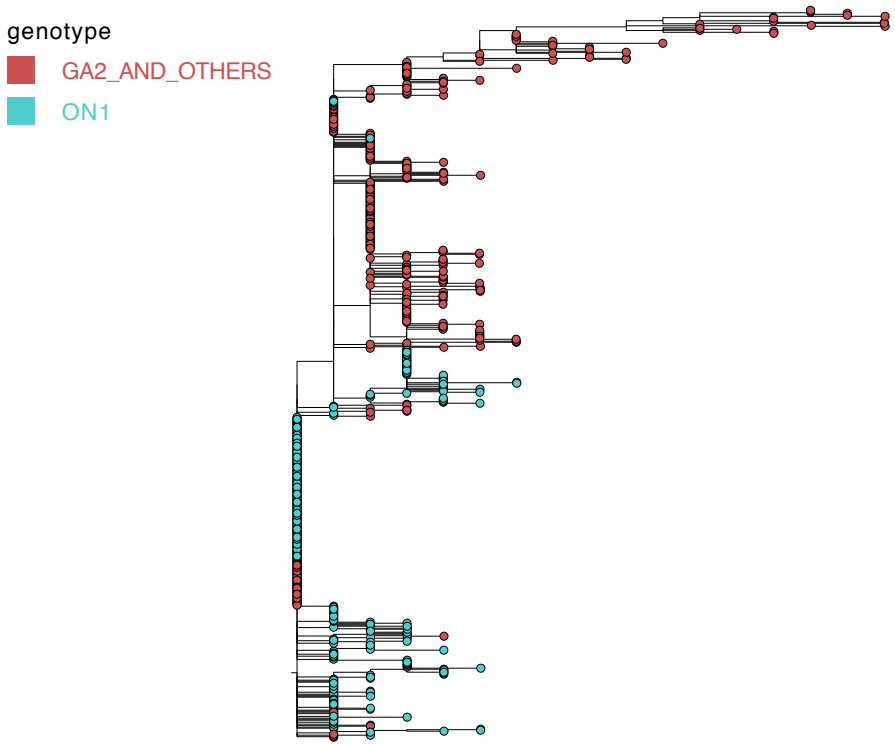
RSV-A SH-gene



RSV-A M-gene

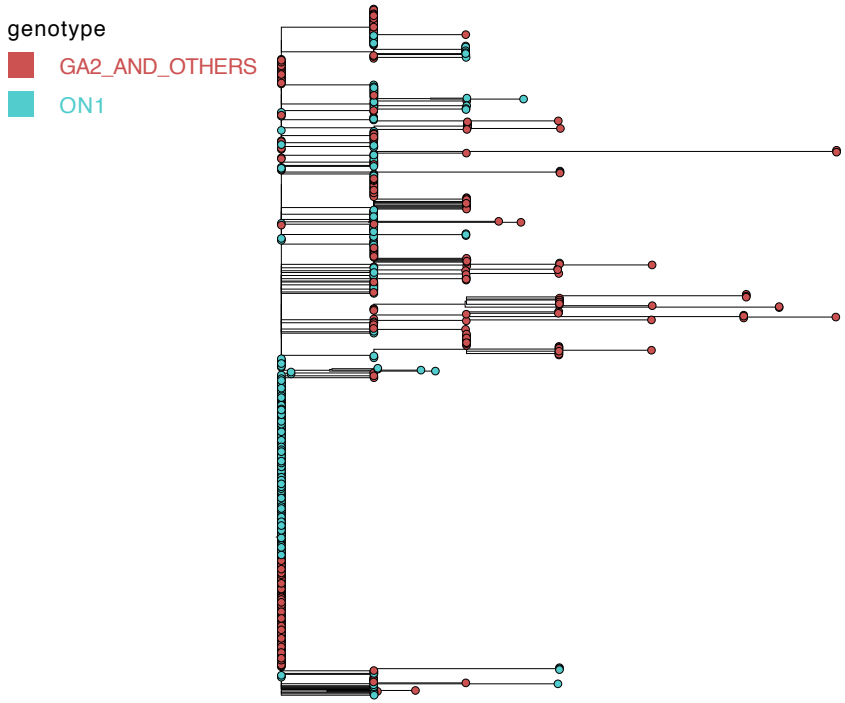


RSV-A M2_1-gene

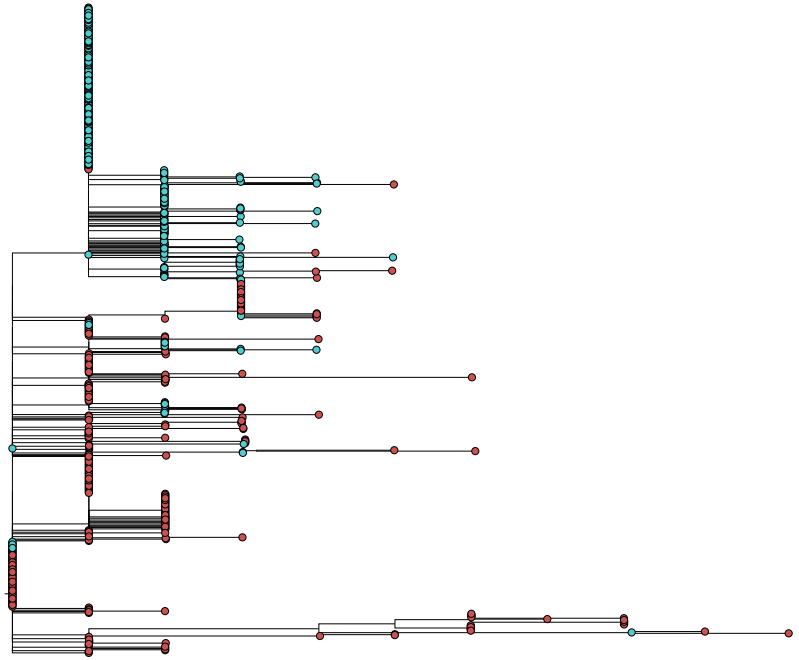


RSV-A M2_2-gene

RSV-A NS1-gene



RSV-A NS2-gene



7.11 Sites under selective pressure

KEY:
* Positive selection; ** Negative Selection; - No site

GENE	SELECTION TEST/ANALYSIS METHOD						
	FUBAR	***SLAC***	***FEL***	***MEME***	***Positive selection by all methods***	***BUSTED***	***aBSREL***
NS1	22**, 23**, 44**, 48**, 67**, 71**, 76**, 94**, 139**	-	22**, 23**, 44**, 71**, 76**, 94**	-	-	Yes	None
NS2	15*, 22**, 70**	22**	19**, 20**, 22**, 54**, 66**, 72**, 96**	-	-	None	None
N	30**, 35**, 49**, 51**, 86**, 100**, 137**, 147**, 153**, 170**, 172**, 177**, 185**, 186**, 205**, 243**, 247**, 253**, 259**, 264**, 267**, 268**, 279**, 316**, 326**, 335**, 350**, 351**, 372**	51**, 153**, 186**, 267**, 259**	30**, 45**, 49**, 51**, 86**, 100**, 137**, 147**, 153**, 170**, 172**, 177**, 185**, 186**, 187**, 205**, 243**, 247**, 253**, 259**, 264**, 267**, 268**, 279**, 326**, 335**, 345**, 350**, 351**, 360**, 372**, 386**	-	-	None	None
P	49**, 68**, 69**, 77**, 98**, 99**, 103**, 129**, 155**, 157**, 163**, 171**, 182**, 184**, 225**	98**, 103**, 129**, 155**, 171**, 225**	39**, 49**, 68**, 69**, 77**, 98**, 99**, 103**, 129**, 155**, 157**, 163**, 171**, 182**, 184**, 194**, 212**, 225**, 240**	-	-	None	None
M	6**, 37**, 41**, 51**, 82**, 92**, 122**, 148**, 157**, 165**, 174**, 188**, 196**, 198**, 208**, 210**, 218**, 250**	6**, 196**, 208**	6**, 37**, 41**, 51**, 82**, 92**, 122**, 148**, 157**, 165**, 188**, 196**, 198**, 208**, 210**, 218**, 250**	-	-	Yes	None
SH	42**	42**	19**, 32**, 42**	-	-	None	None
G	20**, 23**, 80**, 143**, 201*, 202**, 227**, 228**, 239**, 250*, 251*, 310*, 319**	80**, 143**, 227**, 239**, 284*, 310*, 319**	14**, 20**, 23**, 80**, 143**, 176**, 186**, 201*, 202**, 226**, 227**, 228**, 230**, 239**, 251*, 265**, 273*, 289**, 297*, 310*, 319**	73*, 201*, 251*, 273*, 310*	310	None	None
F	6**, 12**, 22**, 26**, 60**, 73**, 100**, 110**, 111**, 112**, 148**, 149**, 154**, 159**, 172**, 182**, 191**, 196**, 215**, 223**, 226**, 259**, 293**, 308**, 348**, 360**, 365**, 405**, 411**, 415**, 422**, 457**, 504**, 528**, 534**, 543**, 555**, 560**, 568**	22**, 73**, 159**, 223**, 405**, 422**, 457**, 534**, 560**, 568**	6**, 22**, 26**, 60**, 73**, 100**, 110**, 112**, 148**, 149**, 154**, 159**, 172**, 182**, 196**, 215**, 223**, 226**, 259**, 293**, 308**, 348**, 360**, 405**, 411**, 415**, 422**, 457**, 504**, 528**, 534**, 543**, 555**, 560**, 568**	-	-	None	None
M2-1	80**, 103**, 120**, 143**, 150**, 186**	103**, 143**, 186**	103**, 125**, 143**, 150**, 164**, 186**	-	-	None	None
M2-2	2**	-	2**, 8**	-	-	None	None
L	22**, 28**, 29**, 30**, 34**, 43**, 73**, 78**, 82**, 99**, 102**, 111**, 114**, 119**, 121**, 152**, 153**, 154**, 176**, 185**, 200**, 210**,	29**, 99**, 121**, 176**, 287**, 392**, 444**, 475**, 533**,	22**, 28**, 29**, 30**, 34**, 43**, 78**, 82**, 99**, 114**, 119**, 121**, 153**, 176**, 200**, 212**, 213**, 221**, 271**, 281**, 334**, 336**,	2030*, 2122*	-	None	None

	212**, 213**, 218**, 220**, 221**, 258**, 271**, 276**, 280**, 281**, 293**, 296**, 314**, 324**, 325**, 334**, 336**, 341**, 349**, 361**, 367**, 368**, 369**, 372**, 375**, 392**, 404**, 405**, 413**, 427**, 436**, 439**, 444**, 446**, 448**, 462**, 463**, 475**, 487**, 488**, 507**, 518**, 526**, 528**, 533**, 540**, 544**, 559**, 563**, 572**, 581**, 591**, 594**, 609**, 610**, 613**, 617**, 618**, 639**, 642**, 648**, 650**, 662**, 689**, 696**, 717**, 719**, 731**, 735**, 737**, 749**, 757**, 758**, 760**, 769**, 770**, 777**, 778**, 790**, 792**, 798**, 808**, 825**, 826**, 831**, 837**, 841**, 846**, 862**, 863**, 868**, 872**, 876**, 878**, 883**, 891**, 892**, 903**, 914**, 933**, 938**, 954**, 961**, 963**, 1004**, 1006**, 1007**, 1016**, 1025**, 1029**, 1033**, 1034**, 1035**, 1042**, 1045**, 1047**, 1049**, 1050**, 1054**, 1074**, 1082**, 1093**, 1094**, 1106**, 1129**, 1135**, 1140**, 1144**, 1145**, 1178**, 1193**, 1197**, 1199**, 1214**, 1236**, 1274**, 1282**, 1286**, 1304**, 1305**, 1310**, 1328**, 1340**, 1345**, 1351**, 1358**, 1362**, 1370**, 1374**, 1376**, 1381**, 1392**, 1397**, 1401**, 1402**, 1408**, 1409**, 1411**, 1414**, 1415**, 1427**, 1435**, 1436**, 1455**, 1483**, 1487**, 1488**, 1496**, 1510**, 1513**, 1522**, 1533**, 1534**, 1542**, 1544**, 1545**, 1546**, 1554**, 1555**, 1561**, 1566**, 1569**, 1576**, 1582**, 1592**, 1593**, 1606**, 1627**, 1638**, 1645**, 1647**, 1667**, 1674**, 1679**, 1681**, 1684**, 1695**, 1696**, 1700**, 1708**, 1712**, 1715**, 1728**, 1735**, 1739**, 1749**, 1755**, 1762**, 1771**, 1773**, 1786**, 1787**, 1799**, 1801**, 1813**, 1819**, 1824**, 1853**, 1868**, 1894**, 1896**, 1900**, 1909**, 1911**, 1915**, 1927**, 1931**, 1944**, 1956**, 1959**, 1961**, 1980**, 2019**, 2024**, 2033**, 2042**, 2060**, 2061**, 2065**, 2075**, 2077**, 2084**, 2095**, 2107**, 2123**, 2124**, 2135**, 2150**	790**, 876**, 954**, 1054**, 1214**, 1297**, 1411**, 1653**, 1773**	349**, 368**, 392**, 427**, 436**, 439**, 444**, 446**, 462**, 463**, 475**, 507**, 518**, 528**, 533**, 544**, 572**, 581**, 591**, 594**, 617**, 639**, 648**, 662**, 717**, 737**, 749**, 758**, 760**, 770**, 778**, 790**, 808**, 825**, 826**, 837**, 846**, 862**, 868**, 872**, 876**, 878**, 883**, 891**, 914**, 938**, 954**, 963**, 1006**, 1016**, 1025**, 1033**, 1034**, 1035**, 1045**, 1050**, 1054**, 1106**, 1140**, 1144**, 1145**, 1193**, 1199**, 1214**, 1236**, 1274**, 1286**, 1340**, 1345**, 1351**, 1358**, 1376**, 1381**, 1397**, 1402**, 1409**, 1411**, 1455**, 1483**, 1487**, 1488**, 1513**, 1522**, 1533**, 1544**, 1545**, 1546**, 1561**, 1566**, 1582**, 1592**, 1627**, 1638**, 1647**, 1674**, 1684**, 1695**, 1700**, 1715**, 1728**, 1735**, 1771**, 1773**, 1786**, 1787**, 1813**, 1819**, 1824**, 1853**, 1868**, 1896**, 1900**, 1909**, 1911**, 1927**, 1931**, 1956**, 2019**, 2042**, 2061**, 2065**, 2077**, 2084**, 2123**, 2124**, 2135**, 2150**				

7.12 Global sampling of RSV-B full G gene and WGS sequences for phylogeographic analysis

G			*WGS*		
Country	***No. of samples***	***Entropy***	***Country***	***No. of samples***	***Entropy***
Argentina	34	-0.1302	Argentina	3	-0.0325
Belgium	3	-0.0202	Belgium	4	-0.0408
Brazil	3	-0.0202	Brazil	3	-0.0325
China	91	-0.2414	China	1	-0.0132
Cuba	38	-0.1405	Germany	3	-0.0325
Germany	3	-0.0202	Italy	7	-0.0631
India	39	-0.1430	Jordan	26	-0.1610
Iran	4	-0.0256	Kenya	16	-0.1158
Italy	7	-0.0400	Mexico	5	-0.0487
Japan	1	-0.0080	New_Zealand	41	-0.2139
Jordan	23	-0.0989	Netherlands	30	-0.1766
Kenya	15	-0.0721	Peru	24	-0.1528
Latvia	3	-0.0202	S. Africa	1	-0.0132
Mexico	5	-0.0306	S. Korea	2	-0.0234
Netherlands	30	-0.1194	UK	29	-0.1728
New_Zealand	29	-0.1166	USA	255	-0.3299
Panama	15	-0.0721	Vietnam	16	-0.1158
Paraguay	1	-0.0080	***Total***	***466***	***1.7383***
Peru	19	-0.0860			
Philippines	29	-0.1166			
Qatar	2	-0.0144			
S. Africa	84	-0.2309			
S. Arabia	2	-0.0144			
S. Korea	36	-0.1354			
Spain	53	-0.1749			
Thailand	27	-0.1109			
UK	29	-0.1166			
USA	196	-0.3401			
Uruguay	1	-0.0080			
Vietnam	14	-0.0685			
Total	***836***	***2.7442***			

G			*WGS*		
Continent	***No. of samples***	***Entropy***	***Continent***	***No. of samples***	***Entropy***
S. America	58	-0.1851	S. America	30	-0.1766
N. America	254	-0.3619	N. America	260	-0.3256
Africa	99	-0.2527	Africa	17	-0.1208
Asia	268	-0.3647	Asia	45	-0.2257
Europe	128	-0.2873	Europe	73	-0.2904
Australia and Oceania	29	-0.1166	Australia and Oceania	41	-0.2139
Total	***836***	***1.568339825***	***Total***	***466***	***1.352903277***

Hemisphere	***No. of samples***	***Entropy***	***Hemisphere***	***No. of samples***	***Entropy***
Southern	63	-0.1948	Southern	44	-0.2228
Northern	452	-0.3325	Northern	331	-0.2430
Tropics	321	-0.3675	Tropics	91	-0.3190
Total	***836***	***0.894855367***	***Total***	***466***	***0.784757005***

7.13 Global sampling of RSV-A full G gene and WGS sequences for phylogeographic analysis

G			WGS		
Country	**No. of samples**	**Entropy**	**Country**	**No. of samples**	**Entropy**
Australia	1	-0.0052	Australia	1	-0.0091
Belgium	2	-0.0093	Belgium	2	-0.0162
Brazil	2	-0.0093	Brazil	2	-0.0162
Canada	9	-0.0324	China	8	-0.0496
China	287	-0.3247	Germany	2	-0.0162
Cuba	46	-0.1121	Hong Kong	4	-0.0286
Germany	2	-0.0093	India	2	-0.0162
Hong Kong	3	-0.0132	Italy	2	-0.0162
India	22	-0.0652	Jordan	38	-0.1541
Iran	3	-0.0132	Kenya	249	-0.3669
Italy	2	-0.0093	Lebanon	1	-0.0091
Japan	5	-0.0201	Mexico	2	-0.0162
Jordan	38	-0.0978	Netherlands	18	-0.0915
Kenya	257	-0.3110	New Zealand	50	-0.1839
Latvia	10	-0.0353	Peru	78	-0.2393
Lebanon	1	-0.0052	Philippines	12	-0.0677
Mexico	9	-0.0324	S. Africa	2	-0.0162
New Zealand	50	-0.1189	S. Korea	1	-0.0091
Netherlands	19	-0.0583	Taiwan	1	-0.0091
Panama	2	-0.0093	USA	220	-0.3616
Paraguay	10	-0.0353	Vietnam	33	-0.1402
Peru	79	-0.1621	***Total***	***728***	***1.8331***
Philippines	18	-0.0559			
Qatar	4	-0.0167			
S. Africa	48	-0.1155			
S. Arabia	11	-0.0380			
S. Korea	19	-0.0583			
Spain	149	-0.2382			
Taiwan	1	-0.0052			
Thailand	28	-0.0782			
USA	229	-0.2960			
Uruguay	3	-0.0132			
Vietnam	33	-0.0882			
Total	***1402***	***2.4923***			

G			*WGS*		
Continent	***No. of samples***	***Entropy***	***Continent***	***No. of samples***	***Entropy***
S. America	94	-0.1812	S. America	80	-0.2427
N. America	295	-0.3280	N. America	222	-0.3622
Africa	305	-0.3318	Africa	251	-0.3671
Asia	473	-0.3666	Asia	100	-0.2727
Europe	184	-0.2665	Europe	24	-0.1125
Australia and Oceania	51	-0.1205	Australia and Oceania	51	-0.1862
Total	***1402***	***1.5946***	***Total***	***728***	***1.5434***

Hemisphere	***No. of samples***	***Entropy***	***Hemisphere***	***No. of samples***	***Entropy***
Southern	51	-0.1205	Southern	51	-0.1862
Northern	733	-0.3391	Northern	253	-0.3673
Tropics	618	-0.3611	Tropics	424	-0.3148
Total	***1402***	***0.8207***	***Total***	***728***	***0.8684***

7.14 Bayes Factor and Posterior Probability support for RSV transition rates between discrete locations in Kenya

Only the transition rates with BF>3 support are shown. In bold are transitions with posterior support of >0.9.

RSV-A:

From	To	Bayes Factor	Posterior Probability
Kilifi	***Siaya***	***15370.18***	***1.00***
Kilifi	***Kakuma***	***15370.18***	***1.00***
Kilifi	***Dadaab***	***3839.34***	***1.00***
Dadaab	***Kilifi***	***849.87***	***1.00***
Siaya	***Nairobi***	***491.68***	***0.99***
Kakuma	***Siaya***	***94.28***	***0.96***
Siaya	***Kakuma***	***39.66***	***0.90***
Kilifi	Kisumu	6.25	0.59
Siaya	Kisumu	4.79	0.53
Kilifi	Nairobi	3.11	0.42

RSV-B:

From	To	Bayes Factor	Posterior Probability
Kilifi	***Nairobi***	***5870.76***	***1.00***
Kilifi	***Siaya***	***5870.76***	***1.00***
Kilifi	***Dadaab***	***5870.76***	***1.00***
Siaya	***Kakuma***	***5870.76***	***1.00***
Dadaab	***Kakuma***	***151.32***	***0.98***
Kilifi	***Kakuma***	***61.29***	***0.95***
Dadaab	Siaya	27.33	0.89
Siaya	Kilifi	20.52	0.86
Kakuma	Kilifi	17.42	0.84